Language Variation and Change (2023), **35**, 223–245
doi:10.1017/S0954394523000121

CAMBRIDGE UNIVERSITY PRESS

Why do we say *them* when we know it should be *they*? Twitter as a resource for investigating nonstandard syntactic variation in The Netherlands

Stefan Grondelaers[1], Roeland van Hout[2], Hans van Halteren[2], and Esther Veerbeek[2]

[1]Royal Netherlands Academy of Arts and Sciences, Meertens Institute Amsterdam, The Netherlands and [2]Radboud University Nijmegen, The Netherlands
Corresponding author: Stefan Grondelaers E-mail: stef.grondelaers@meertens.knaw.nl

Abstract
Two Twitter-based corpus studies are reported to account for the increasing preference in The Netherlands for the stigmatized subject use of the object pronoun *hun* 'them.' Twitter data were collected to obtain a sufficient number of *hun*-tokens, but also to investigate the validity of two hypotheses on the preference for *hun*, this is, that subject-*hun* is a contrast profiler which thrives in contexts of evaluation and qualification, and that subject-*hun* is propelled by its dynamic social meaning, being a tool for nonposh and streetwise self-stylization. Although the latter is not normally a predictor included in regression analyses of constructional choice, it turns out that expressively spruced up tweets with vivid contrast profiling are the prime biotope of subject-*hun*. Along the way, this paper reviews the potential of Twitter data for the reconciliation of macro-big-data analysis with micro-sociolinguistic focus, but it also reports and attempts to remedy three concerns.

Keywords: syntactic variation; pronoun diffusion; Twitter; prestige; stylization

This paper tests the suitability of Twitter materials for the investigation of the stigmatized subject use of the object pronoun *hun* 'them' in Netherlandic Dutch, as in (2) versus (1):

1. *Als je zo speelt krijgen **zij** natuurlijk altijd kansen.*
   'When you play like that **they** will of course always get chances'
2. *Als je zo speelt krijgen **hun** natuurlijk altijd kansen.*
   'When you play like that **them** will of course always get chances'
   (Van Hout, 2003:277)

This subject use of the object pronoun *hun* (henceforth subject-*hun*) is the most notorious Netherlandic diffusion of the past decades. It was first observed in Vor der Hake (1911), but it lay dormant in the grammar of Dutch before receiving a major impetus in the past four decades. In spite of the fact that few linguistic

diffusions have been so hysterically mediatized in The Netherlands as subject-*hun* (Grondelaers, Van Gent, & Van Hout, 2022), there is wide agreement (see, among many others, Grondelaers & Van Hout, 2021; Van Bergen, Stoop, Vogels, & De Hoop, 2011; Van Hout, 2003) that subject-*hun* is leaving its original "habitat" of young, lowly educated, informal, and unscripted, and that it is rapidly conquering Dutch, also in more formal contexts.

Yet, due to the scarcity of suitable data resources, there is almost no empirical evidence that documents the conditioning of subject-*hun*. The largest available corpus of Netherlandic Dutch, the Twente News Corpus (Ordelman, De Jong, Van Hessen, & Hondorp, 2007), totals half a billion words but consists exclusively of print newspapers that shun prescriptively deviant usage. The Spoken Dutch Corpus (Oostdijk, 2003) contains dialogues that are sufficiently informal to feature subject-*hun*, but it is much smaller (nine million words), and the absolute frequency of subject-*hun* is low ($n = 213$) and disproportionate in comparison with standard subject-*zij* ($n = 1048$). In this paper, we take the quest for the optimum data source to study nonstandard variation one step further and propose Twitter data for the investigation of subject-*hun*. Twitter was launched in 2006 and is available in enormous quantities, featuring language use with many characteristics of orality (Androutsopoulos, 2011) in which prescriptivism plays a lesser role. For this reason, Twitter materials have been found to be eminently suited as "supplementary data for investigating non-frequent, non-canonical phenomena in spoken language" (Rehbein, 2014:20). We use Twitter materials to obtain a number of subject-*hun* tokens sufficient to investigate grammar-internal accounts of the preference for *hun*. But we also rely on tweets to follow up on investigations that have found that accelerating features are frequently linked to associations of streetwise urban cool and toughness (see Stuart-Smith, Pryce, Timmins, & Gunter, 2013, for an overview). Grondelaers, Van Gent, and Van Hout (2022) provided evidence that the propagation of subject-*hun* is also linked to urban cool associations and dynamic prestige, but the data on this correlation were experimentally elicited. If we want to demonstrate that the preference for *hun* is related to dynamic prestige associations, we need to compare these prestige propellers to grammar-internal and -external predictors in one encompassing regression analysis. Such an integrated analysis presupposes that we can infer prestige associations from production data, and for this ambition too Twitter fosters possibilities.

Against the backdrop of our ambition to provide an integrated account of the preference for subject-*hun*, we investigate the appropriateness of Twitter materials for sociolinguistic analysis. There are three central concerns in this respect. To begin with, does an eminently vernacular variant survive the amount of attention paid to the conscious mode of writing? Second, to what extent is the availability of a larger number of nonstandard tokens in tweets offset by the impossibility to code for specific demographic predictors? And third, to what extent does the exodus of young tweeters to other social media outlets around 2014 impact the frequency and conditioning of nonstandard variants?

This is how we will proceed. In the next sections, we zoom in on prior work pertaining to subject-*hun* and review previous studies relying in some way on Twitter data. We then outline our research questions and hypotheses. In a first corpus-based study, we test the validity of grammar-internal and prestige-related predictors in a

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

large Twitter dataset from 2014. The remaining concerns are addressed in a smaller-scale study with a diachronic dimension, more demographic predictors, and an independent contrast implementation. After that, we interpret our findings in terms of the research questions, and the final section brings in some theoretical consequences of our data.

Prior analyses of subject-*hun*

A number of prior production studies have suggested grammar-internal accounts for subject-*hun*. Van Bergen et al. (2011) proposed that *hun* has a more specific meaning than standard *zij* and the reduced standard form *ze*, because it exclusively refers to animate and especially human referents, while *zij* and especially reduced *ze* can also denote nonanimate entities. In a corpus analysis, they found that *hun* never refers to inanimate entities, while *zij* occasionally does (albeit in only three out of five hundred random tokens).

An alternative hypothesis proposed in Grondelaers and Van Hout (2021) is that *hun* is better suited to encode "vivid (negative) contrasts," as in *Wij zijn Ajax, wegwezen met al de rest. Hun horen het in hun broek te doen* 'We are Ajax, to hell with all the rest. Them are supposed to shit their pants.' The reason for this is that *hun* typically bears sentence stress and that it is phonetically unreducible on account of its consonant in final position; standard *zij*, by contrast, can be reduced when it does not bear stress. In order to test both Van Bergen et al.'s (2011) animacy hypothesis and their own contrast hypothesis, Grondelaers and Van Hout (2021) extracted 2,449 sentences with either subject-*ze*, -*zij*, or -*hun* from a corpus of 125 hours of Netherlandic Dutch television. Their main finding was that unreduced pronouns (*hun* and *zij*) were typically avoided when reference with the statistically dominant (91.38%) reduced form *ze* will do; as a consequence, they argued that subject-*ze* should be removed from the envelope of variation. Regression analysis furthermore demonstrated that *hun* is preferred in reality TV shows (like *Big Brother*), in contexts where the pronoun bears stress, and when the pronoun refers to a third party that is negatively contrasted with the speaker group. There was no evidence for the animacy hypothesis: neither subject-*hun* nor subject-*zij* ever referred to inanimate entities. In a second study, Grondelaers and Van Hout (2021) reanalyzed Van Bergen et al.'s (2011) original dataset but also coded predictors in function of the contrast hypothesis and included the demographic and situational variables available in the Spoken Dutch Corpus. It was shown, once more, that contrast profiling was the most important internal predictor and that the animacy effect was in fact an *individuation* effect: subject-*hun* shuns the type of group reference illustrated in (3) and prefers individuated reference, as in (4):

3. This is the housing department: they will tell you where to go.
4. These are the housing officers: they will tell you where to go.

The external predictors, however, were by far the most important for model fit: at the beginning of the twenty-first century, it was predominantly female, younger, and lower educated people who preferred *hun*.

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

In addition to the contrast profiling benefit of subject-*hun*, Grondelaers et al. (2022) propose a social meaning explanation for its diffusion, in line with a growing number of studies that demonstrate that accelerating features are linked with associations of streetwise cool and urban toughness (see Foulkes & Docherty, 1999; Grondelaers & Marzo, 2023; Sneller & Roberts, 2018; Stuart-Smith et al., 2013). Respondents were asked to associate spoken stimuli with two sets of pictures, either representations of traditional prestige associated with standard usage (e.g., an antique Chesterfield couch or a symphonic concert venue), or representations of urban cool associated with nonstandard usage (such as a flashy sushi restaurant or an internet company in a recycled factory hall). Grondelaers et al. (2022) found that whereas the standard form *zij* was upgraded on both prestige representations, subject-*hun* was upgraded only on the modern prestige representations.

Linguistics on Twitter

Twitter was launched in 2006 as a microblogging platform for messages up to 140 characters (280 since 2017). Twitter shares with other computer-mediated communication (CMC) platforms a number of characteristics that can be encapsulated in two crucial principles (Androutsopoulos, 2011:149).

The *conceptual orality* principle (see also Hilte, Vandekerckhove, & Daelemans, 2018) pertains to netspeak's approximation of casual speech features, instantiated in the omnipresence of nonstandard orthography that is either the result of error, or—more interestingly—of expressive or indexical resourcefulness (Coats, 2016:188). Crucially, Twitter shares with authentic vernacular speech the presence of phonetic, lexical, and morphosyntactic cues revealing demographic properties of tweeters. Twitter distributions of these features, moreover, pattern so well with traditionally observed distributions that they are eminently suited for probing regional patterns that would otherwise be too laborious to investigate. For example, Jones (2015) demonstrated that orthographically represented features of AAVE on Twitter pattern with the geographic spread of the Black population in the United States, Brown (2016) investigated the distribution of a Spanish construction on a much wider geographic scale than before, and Haddican and Johnson (2012) found that respondents from the UK and Ireland favored discontinuous particle verb orders (*she cut the melon open*), while US and Canadian participants preferred the continuous order.

The *expressive compensation* principle pertains to the ubiquity in CMC of strategies like lengthening (*coooool*), capitalization (*COOL*), excessive punctuation (*cool!!!*), intensification (*fckng cool*), and emoticons and emojis, which are used to "compensate" for the absence in a written medium of expressive features like intonation and facial or manual gestures. The most studied expressive resource, expressive lengthening, which is also dubbed "word lengthening" (Brody & Diakopoulos, 2011) or "flooding" (Hilte et al., 2018), mimics dynamic intonation and prosody and conveys positive and negative emotions (De Decker & Vandekerckhove, 2017), including sadness, happiness, disappointment, doubt, and sarcasm (Parkins, 2012). There is widespread agreement that lengthening is a (very) young expressive resource. While Verheijen (2018) found that Dutch adolescents used lengthening significantly more often than young adults, De Decker and Vandekerckhove (2017) found that younger

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Flemish adolescents (13-16) produced more lengthened words than an older group (17-20).

Nevertheless, three concerns remain for Twitter-based studies. The first pertains to a number of remarkable changes in the frequency and demography of (Dutch) tweets around the year 2014, which potentially compromise Twitter-based diachronic studies. Sanders (2023) builds on the TwiNL-corpus we use in the upcoming studies, a gigantic collection of Dutch tweets extracted on the basis of a keyword-based stream that was activated in December 2010 (Tjong Kim Sang & Van den Bosch, 2013) and currently totals more than eight billion tweets. Sanders (2023:3) noticed that between 2013 and 2014, the number of tweets in the TwiNL-corpus just about halved, going from almost eight hundred million in 2013 to less than four hundred million in 2014. Since nothing changed in the way the tweets were pulled according to the TwiNL-compilers (Sanders 2023:3, referring to personal communication with TwiNL-compiler Tjong Kim Sang), they attributed this halving to the fact that many young tweeters left Twitter in 2013-2014 for other social media platforms like Instagram, Snapchat, and (from 2016) also TikTok (an explanation seconded in Sanders, 2023:3). While we have not found any scholarly corroboration of this defection, it is well documented on popular news sites, though the change is situated somewhat later: since early 2016, according to Suciu (2022), "[Twitter] has seen a steady decline of those in Generation Z as well as Millennials. According to a study from YPulse conducted earlier this year (…) Twitter's decline among younger users is part of an ongoing trend. It found that in March 2016, 51% of Millennials and 42% of Gen Z respondents said they used the platform, while that number has fallen to 32% and 28% respectively." In The Netherlands, Turpijn, Kneefel, and Van der Veer (2015) found that use of Twitter rapidly declined after 2014 and that the change was especially noticeable in the youngest demographic.

A second concern pertains to the general absence of demographic information on tweeters. Only a small proportion of tweets—3% to 4%—is stored with geolocation coordinates; independent information on tweeters' gender, educational background, or socioeconomic class is even scarcer. In addition, as we will see in Study 1, Twitter is a veritable "Pandora's Box," featuring a bewildering diversity of text genres and interactional settings that have not, to our knowledge, been fully appreciated as conditioning predictors.

Third, the coding of language-internal constraints is not a straightforward enterprise either with Twitter data. Recall that tweets are restricted in length (140 characters up to 2017), and that they do not represent running discourse in the shape of linearly organized text or dialogue. These restrictions endanger responsible coding for complex predictors like the contrast implementation in Grondelaers and Van Hout (2021), which requires longer stretches of running discourse. In addition, the main advantage of Twitter datasets, their large size, precludes the possibility of extensive hand-coding. Many colleagues who have availed themselves of extensive Twitter datasets typically use computational tools to code them semiautomatically. A case in point is Bohmann's (2016) investigation of the growing inclination of *because* to take adjectives (*Early morning gym because fat*) or nouns (*He didn't come home today because work*) as complements. In view of the recent emergence of this innovation (mid-1990s), it is still a low-frequency phenomenon, and almost no predictors

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

have been proposed for it, except that it is "exceptionally bloggy and aggressively casual and implicitly ironic" (Garber, 2013, quoted on p. 157) and that it has "a snappy, jocular feel to it" (Carey, 2013, quoted on p. 159). While all these associations pattern well with the urban cool traits that have been proposed as the social meaning drivers of a number of rapid innovations (including subject-*hun*), Bohman (2016:159) declined to code for them "since it is near impossible to model irony or humor quantitatively." Building on automated scripts, he did code a set of 12,751 geolocated tweets for fifteen predictor variables, including measures of colloquialization (as a proxy for the casual and bloggy associations), and for typical CMC features such as hashtags, @-mentions, URLs, and emoticons. Crucially, Bohmann did not find a correlation between *because X* and overall informality or colloquialness. He attributed this failure to the possibility that *because X* does not code "the casualness of unmonitored quotidian talk, but a studied, consciously constructed one that is exploited as a poetic device. In other words, *because X* is perhaps not so much an indicator of a generally casual style as a resource that is exploited in the *stylization* of casualness (Coupland 2001)" (2016:175).

Following up on the crucial issue of coding for stylized casualness, Grondelaers and Marzo (2023) argued that Twitter's expressive compensation strategies can be used to operationalize young, cool, and nonposh self-profiling. They first carried out a speaker evaluation experiment into the dynamic prestige boosts of an ethnolectal diffusion in Flanders. They then switched perspective from dynamic prestige as a *perception* phenomenon, to the production cues in the CMC-toolbox that tweeters can employ to *stylize* themselves as dynamic, nonposh, and streetwise, including lengthening, capitalization, intensifiers and interjections, and excessive punctuation marking. In a subsequent Twitter-based corpus study, they found that expressiveness, as measured on these features, was the second most important predictor of the investigated ethnolectal forms.

Research questions and hypotheses

In the upcoming study, we follow up on the reported research on subject-*hun* by testing two hypotheses:

Hypothesis 1: subject-*hun* is internally constrained to contexts of contrastive evaluation and qualification.

Hypothesis 2: subject-*hun* is associated with dynamic social meanings (viz. associations of cool, nonposh, streetwise, tough, cheeky, etc.), and, in this capacity, it is a tool for nonposh and lively self-stylization.

In order to test these hypotheses, however, we first have to answer the following research questions pertaining to the potential but also the problems of Twitter as a data resource for the study of morphosyntactic change:

RQ1 Can Twitter help us solve the data problem encountered in earlier studies of nonstandard preferences?

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

(a) Does the more conscious mode of writing bar the occurrence of stigmatized vernacular forms like subject-*hun*?
(b) Does Twitter allow us to compile a large dataset that features a sufficient amount of subject-*hun* tokens?

RQ2 Are Twitter data suitable for sociolinguistic study?

(a) Does the length condition on tweets (max. 140 characters up to 2017) allow us to code for complex semantic predictors like contrast?
(b) How do we get access to external predictors that are not independently available in Twitter materials?

RQ3 How does the alleged defection of younger tweeters around 2014 impact the distribution and the modeling of subject-*hun*?

Study 1 specifically tackles RQ's 1 and 2, building on a large corpus of tweets from 2014, specifically compiled to address the data scarcity problem. Study 2 relies on a smaller, but more richly annotated corpus of tweets spanning the period between 2011 and 2019, compiled to address RQ2b and especially RQ3.

Study 1. Investigating the frequency and conditioning of subject-*hun*

Materials

Our dataset was extracted in April 2014 from the TwiNL-corpus. In light of the absence of performant POS-taggers for Dutch, and in order to guarantee that our query selected tweets with the intended use of subject-*hun*, we extracted tweets in which *hun* or *zij* preceded twenty-three different plural verb forms[1] that frequently collocate with these personal pronouns (as revealed by a prior corpus analysis). In addition, we limited our search to the one thousand most frequent hashtags in the corpus in order to obtain some contextual and topical background to the tweets. In spite of these restrictions, we ended up with a much larger and more balanced dataset than the Television corpus in Grondelaers and Van Hout (2021): while subject-*hun* in the latter was statistically marginal (representing 2.08% of all tokens when *ze* was included in the envelope and going up to 24.17% when *ze* was excluded), our Twitter dataset initially featured 7,112 tweets with standard *zij* (48.66%) and 7,504 with nonstandard *hun* (51.34%).

All tweets were stored in the database with their tweet ID, the screen name of the tweeter, and the main verb—which always follows the pronoun—in separate fields.

Predictors and bivariate analyses

Tweets were hand-coded for six predictors selected in function of the animacy hypothesis, the contrast hypotheses, and the social meaning(s) alleged to correlate with subject-*hun*'s diffusion. The individual effect of these predictors will be tested in separate chi-squared tests reported in summary Table 1.

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

In order to test the animacy hypothesis, we initially applied the four-level taxonomy from Grondelaers and Van Hout (2021) and classified pronoun referents in "human individuals" (… *the ladies$_i$, they$_i$*) versus "human collective" (… *consulted the welfare department$_i$, they$_i$ told us*…) versus "nonhuman animal" (…*heard the dogs$_i$, they$_i$*…) versus "nonanimate" (… *my feet$_i$, they$_i$ started to hurt*). The animacy of fifty-two *hun*-antecedents and eighty-one *zij*-antecedents could not be determined, and these 134 tokens were discarded from further analysis. If anything, our data confirm Grondelaers and Van Hout's (2021) conclusion that there is no support for the animacy hypothesis: *zij* only rarely refers to nonanimate entities ($n = 30$; 0.42%), and *hun* does refer to an inanimate entity once:

5. *okeee, woow* ***hun*** *zijn goed die rollschaatsennn. #hgt*
   'okay, wow **them** are good these roller skatesss. #hgt'

The fact that the lexeme *rolschaatsen* 'roller skates' is expressively lengthened on its final -n indicates that it is not an erroneous variant of *rolschaatsers* 'roller skaters.' In all further analyses, we will discard the ninety-six tweets that instantiate the nonhuman levels 3 ($n = 65$) and 4 ($n = 31$), as such low frequencies do not permit closer investigation of other linguistic factors; this leaves us with a dataset of 14,387 tokens.

Coding for the contrast hypothesis requires some inventiveness when working with materials that do not feature running discourse. There is a considerable number of tweets in our dataset that explicitly code Grondelaers and Van Hout's (2021) strongest type of contrast, that is, between the speaker (group) and a third party, as illustrated in (6)-(7):

6. *#ajax calimero gedrag, ingegeven door #cruijff : wij zijn dom en* ***hun*** *zijn rijk?*
   '#ajax calimero behaviour, inspired by #cruijff : we are stupid and **them** are rich?'
7. *wij zoeken gewoon. altijd heelweinig ma superrleuk uit, en* ***hun*** *denken modieus tezyn, en zoeken jaren 50 uit #haha*
   'we try to select. always very little but extremely nice, and **them** believe to be fashionable and look for fifties #haha'

Tweets like (6)-(7) instantiate the template [pronoun + copula + adjective], an unadorned attributive frame that highlights the quality that is the central element of the proposition. In (6), the double use of this construction explicitly foregrounds the contrasting properties *dom* 'stupid' and *rijk* 'rich'; in (7), the evaluative template is somewhat more modal (the hashtag signals the ironic intent), but the we-them opposition is no less obvious.

Examples (8)-(11) are only slightly less contrastive. The tweet in (8) makes no explicit reference to the speaker (group), but the copula construction serves to single out the cultural differences between the Flemish and the Dutch contestants of the survival television program *Expeditie Robinson*. Even the bare templates in (9)-(11)—in which the third party is evaluated rather than explicitly distinguished from the tweeter's own group or some other party—are inherently contrastive for two reasons. The hashtagged reference to some televised talent contest delimits an arena in which positive and

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

negative evaluations of the same content are juxtaposed. In this sense, tweets such as (9)-(11), which comment on the performance of a formation called "b-brave" in the talent contest *X-Factor*, represent an act of evaluation, qualification, categorization, *and* contrasting. In addition, televised talent contests are typically geared toward showcasing the merits of nonprofessional performers, which are more often than not the *peers* of their evaluators rather than acclaimed idols tweeters can look up to. Any sort of evaluation in this context therefore automatically entails an act of contrasting.

8. ***hun*** *zijn allemaal super anders x #expeditierobinson*
   '**them** are all super different x #expeditierobinson'
9. *b-brave de niewe one direction laat me niet lachen man* ***zij*** *zijn fucking slecht #xfactor*
   'b-brave the new one direction, don't make me laugh man **they** are fucking bad #xfactor'
10. *#xfactor b-brave moet echt door* ***hun*** *zijn zo goed ze moeten. winnen* 😊
    '*#xfactor* b-brave should go on to the next round **them** are so good they have to. win'
11. *oeehh hun* ***zijn*** *echt knap #xfactor b-brave*
    'owww **them** are really handsome #xfactor b-brave'

In order to test our proposal that the predicative copula template [third-person plural pronoun + copula + quality] is a contrastive structure par excellence, we classified tweets in our dataset in terms of construction type, distinguishing between the levels [pronoun + copula + adjective] (as in 6-11) and all other construction types.

We also coded our tweets for expressive compensation strategies such as intensifying adverbs (*super* in [7]-[8], *fucking* in [9]), lengthening such as *okeee* and *rollschaatsennn* in (5), interjections (*oehh* in [11]), and CMC-additions such as emojis (as in [10]). Following Grondelaers and Marzo (2023), we regard these strategies as resources tweeters can apply to "tweak" their personality, and to stylize themselves in terms of qualities (young, cool, dramatic, ironic, etc.) that readers can pick up as dimensions of a dynamically prestigious personality that is prone to *hun*-use. The factors intensifiers, interjections, lengthening, and CMC-additions (a category that included emojis and emoticons but also CMC-textisms like *lol* or *yolo*) were coded in terms of presence or absence.

The impact of all the investigated predictors on *hun*-usage is reported in summary Table 1, which diagrams, per predictor, the chi-square statistic, the effect size estimate φ, and the *p*-value, and, for each predictor level, absolute frequencies of *zij* and *hun*, and the relative frequency of *hun*.

All factors in Table 1 are highly significant determinants of *hun* (all $p < .000$), but their φ-coefficients demonstrate that construction type, our proxy for contrast, is the strongest: subject-*hun* specifically thrives in copular constructions ($\varphi = 0.386$). The data also reveal a crucial effect of individuation ($\varphi = 0.276$): when reference to individuated humans is made, *hun* is about three times more frequent (57.3%) than when people in a collective are referred to (17.4%). Among the expressive compensation strategies, intensification clearly is the more important *hun*-booster ($\varphi = 0.332$).

Before we turn to regression analysis to obtain a multivariate picture of *hun*-preference, two crucial observations with respect to our implementation of contrast

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Table 1. Cross-tabulations of subject pronouns *hun* and *zij* by six predictors

	n zij	*n* hun	% hun
Individuation ($\chi 2 = 1097.4$; $\varphi = 0.276$; $p < .000$)			
group	1656	349	17.4
Individuated	5289	7093	57.3
Construction type ($\chi 2 = 2137.6$; $\varphi = 0.386$; $p < .000$)			
non-copular	5525	3107	36.0
copular	1420	4335	75.3
Intensifiers ($\chi 2 = 1584.9$; $\varphi = 0.332$; $p < .000$)			
0	5857	3976	40.4
1+	1088	3466	76.1
Interjections ($\chi 2 = 607.4$; $\varphi = 0.206$; $p < .000$)			
0	6382	5719	47.3
1+	563	1723	75.4
CMC-additions ($\chi 2 = 111.2$; $\varphi = 0.088$; $p < .000$)			
0	6468	6545	50.3
1+	477	897	65.3
Expressive lengthening ($\chi 2 = 419.2$; $\varphi = 0.171$; $p < .000$)			
no	6727	6521	49.2
yes	218	921	80.9

in terms of the copula template should be made. First, there is no one-on-one relation between contrast and construction type: example (2) above instantiates a noncopular construction that is clearly contrastive.

More importantly, it is unclear whether the copular format exclusively indexes contrast, for construction type arguably cross-classifies with situational constraints and a demographic predictor (age) that is a proven *hun*-inducer. Some evidence for the situational and demographic correlates of construction type is presented in Table 2, which ranks hashtags for copular and noncopular constructions in terms of frequency (relative frequencies represent the ratio of a hashtag and the total number of hashtags represented in respectively copular and noncopular tweets in the dataset; it should be recalled at this point that we exclusively extracted hashtagged tweets and that tweets often contain more than one hashtag). Table 2 demonstrates that there is a noticeable connection between the copular format and televised contest shows such as *Holland's Got Talent* (#hgt), *X-Factor* (#xfactor), *So You Think You Can Dance* (#sytycd), *The Voice of Holland* (#tvoh), *The Voice Kids* (#tvk, the junior version of *The Voice of Holland*), *The Ultimate Dance Battle* (#tudb), and, somewhat more traditional, the *Eurovisie Songfestival* (#esf, Eurovision Song Contest). While contest-related tweets (shaded in gray) dominate the ranking for both construction types, they are much more frequent in the copular.

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Table 2. Hashtags ranked per absolute and relative frequency in (non)copular constructions (contest-related hashtags are marked in gray)

	Copular			Noncopular		
Rank	Type	Absolute frequency	Relative frequency	Type	Absolute frequency	Relative frequency
1	#hgt	2313	33.4	#hgt	447	3.8
2	#xfactor	815	11.8	#xfactor	253	2.2
3	#sytycd	496	7.2	#fail	173	1.5
4	#tvoh	299	4.3	#tvoh	157	1.3
5	#rtl4	91	1.3	#ff	144	1.2
6	#tvk	89	1.3	#pvv	105	0.9
7	#mtv	56	0.8	#penw	92	0.8
8	#gtst	53	0.8	#ajax	90	0.8
9	#not	42	0.6	#gtst	90	0.8
10	#haha	37	0.5	#respect	88	0.7
11	#tudb	32	0.5	#sytycd	86	0.7
12	#1	29	0.4	#dtv	77	0.7
13	#love	29	0.4	#vvd	72	0.6
14	#respect	25	0.4	#haha	70	0.6
15	#rtl5	23	0.3	#pvda	65	0.6
16	#lol	22	0.3	#kenjedat	64	0.5
17	#esf	20	0.3	#cda	56	0.5
18	#orpheus	19	0.3	#in	53	0.5
19	#ajax	18	0.3	#jaloers	52	0.4
20	#emidj	18	0.3	#ns	52	0.4

Whereas the seven contest formats tagged in the copular column in Table 2 make up no less than 58.6% of all the hashtag tokens in the copular constructions, the four contests tagged in the noncopular formats represent only 8% of the hashtags in that construction. This goes to show that the copula construction is not only an obvious vehicle for contrast profiling, but that it also indexes specific topics and interactional settings.

These topics and interactional settings, in turn, arguably index other constraints. We may not have direct access to our tweeters' age, but the hashtag rankings in copular and noncopular formats suggest very different media preferences, which arguably correlate with an age distinction. In addition to a predilection for televised contest shows, the left column of Table 2 suggests that it is mainly commercial broadcasters (#rtl4, #rtl5, #mtv) and young programs (#emidj tags the reality show *Echte meisjes in de jungle* 'real girls in the jungle') that are being tagged in copular constructions,

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

whereas hashtags in the right column feature political parties (#pvv, #vvd, #pvda, #cda) and the high-brow talk show *Pauw & Witteman* (#penw).

An indirect quantitative indication that the copular template is associated with younger people is the observation that expressive compensation strategies that have been found to be used much more frequently by younger tweeters (see the Linguistics on Twitter-section), are also used much more frequently in copular templates, as shown in Table 3.

In addition to the fact that the expressive strategies in Table 3 are more frequent in the copula construction, specific examples of the latter like (9)-(11) also "sound" young: the uncritical and unrestrained support for, or disapproval of b-brave (a Dutch boy band) strongly bespeaks an adolescent audience. As a consequence, it is uncertain at this point whether the higher *hun*-proportion in the copular template is a function of the semantic predictor contrast profiling, of the genre televised contest, or of the external constraint age—or of all three.

Table 3. Relative frequencies of expressive compensation strategies as a function of construction type

	Intensifiers	Interjections	CMC-additions	Lengthening
−copula construction	12.6	9.1	8.3	2.8
+copula construction	59.2	25.7	11.4	15.3

Regression analysis

We fitted a logistic mixed effect regression (*R*-package *lm4*, option *glmer* with the *logit* function; *R*-package *sjPlot*) in which we entered as fixed effects individuation, construction type, and an aggregate measure of expressiveness/self-stylization. We miss a generalization if we regard the expressive strategies as separate predictors of *hun*, for they represent resemblant tools with an analogical function which are not, however, implicationally related (to the extent that use of one entails use of the others). For our modeling of subject-*hun*, to be sure, it may not be the *nature* of the expressive strategies that is decisive but rather the number of them. As a consequence, we created a new predictor, *expressiveness*, for which we counted the number of different types of strategies, initially with levels 0 (no expressive strategies) to 4 (for tweets containing intensifiers, interjections, additions, and expressive lengthening). Since cross-tabulating the effect of these levels on pronoun preference demonstrated that the presence of more than two strategies does not noticeably increase *hun*-use, we eventually defined three levels (zero versus one versus twoplus).

Verb ($n = 23$) was included as a random effect. The best model was selected through ANOVA-based comparison of nested models and AICs. In view of the special status of the factor construction type—a potent *hun*-predictor which plausibly co-varies with age—we included all predictors as main effects but also in interaction with construction type.

Figure 1 diagrams the statistical significance and the effect size (odds ratios) of the main effects and interactions included in the best model. Positive odds ratios on the right of the 1-mark on the horizontal odds ratios axis indicate that *hun*-preference increases on account of a main effect or interaction. The positive odds ratio of 6.40

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Figure 1. Model plot (with odds ratios) of logistic mixed effect regression on *hun*-preferences in the full dataset (* $p < .05$, ** $p < .01$, *** $p < .001$).

for expressiveness [twoplus] in Figure 1, for instance, indicates that the odds for *hun* mathematically increase 6.4 times when a tweet contains two or more expressive strategies compared to when a tweet contains no expressive strategies. Negative odds ratios on the left of the 1-mark indicate that *hun*-preference decreases on account of a main effect or interaction; the negative odds of 0.32 for the interaction between construction type and expressiveness [twoplus], for example, reveals that the statistical odds for *hun* decreases $1/0.31 = 3.13$ times when two or more expressive strategies occur in a copular construction type, compared to when they occur in a noncopular format.

In Figure 1, all main and interaction effects are statistically significant. In order of impact, it is especially expressiveness which enhances the preference for *hun*. The lower significance ($p \leq .05$) and odds ratio (2.55), and the larger confidence interval (indexed by the horizontal line that has the odds ratio in its center) for construction type signal uncertainty about the status of that factor as a main effect and force us to interpret it in interaction with the other predictors. The interaction individuation*construction type reveals that reference to individuated persons rather than to a collective (such as *dance group* or *police*) is a much stronger *hun*-inducer in the copula constructions than in the noncopular construction. The interactions between construction type and expressiveness show that the main effect of expressiveness as a *hun*-booster is significantly *reduced* in the copula construction: expressive compensation strategies may be more frequent in copular constructions (as shown in Table 3), but they predominantly enhance *hun*-use in the noncopular formats.

Interpretation

The findings of Study 1 have shown that the conditioning of *hun* is more complicated than hypotheses 1 and 2 (in the Research Questions and Hypotheses section) prepare

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

for. What is the exact relation between contrast/construction type, dynamic prestige/expressive self-stylization, and individuation? Let us provisionally disentangle these factors for both construction types separately.

The copular construction is an evident *hun*-booster in its own right (given the data in Table 1), but recall that it is difficult to pinpoint the exact driving force—contrast or young age—that it represents (we return to this issue in the General Discussion). In the context of televised contest shows like *Holland's Got Talent* or *So You Think You Can Dance*, the evaluated amateur performers are more likely than not the *peers* of young evaluators, and the conceptual proximity between evaluator and "evaluee" makes for the most engaged contrast and, hence, for the highest *hun*-proportions. The interaction between individuation and construction type is easy to explain in this light: it is logical that *hun's* preference for individuated human evaluees instead of abstract collectives should be greater in the copular template favored in tweets commenting on talent shows.

The noncopular group contains a more heterogeneous bag of constructions, topics, and users. In addition, the fact that constructions in this category need not be as intrinsically evaluative or contrastive as the copular template (a reality reflected in the much lower *hun*-proportion in Table 1), does not mean that they cannot, or do not, express contrast. This renders a straightforward interpretation of the odds ratios in Figure 1 hazardous. An important observation is, in any case, the observed predominance in noncopular constructions of the expressive compensation strategies as *hun*-boosters.

Study 1 has demonstrated, in any case, that the preference for subject-*hun* can be modeled fairly successfully along lines that have been proposed in earlier research: both contrast profiling and dynamic prestige (as measured on their proxies of construction type and expressiveness) have been shown to be evident *hun*-boosters. What this study has documented first and foremost, however, is the challenges inherent in using Twitter data for sociolinguistic analysis. Twitter's pivotal advantage as a source of rich datasets featuring standard and nonstandard tokens is crucially offset by the unavailability of crucial predictors like age, gender, and education level, and by the difficulty to code for contrast in a responsible way: we have shown that construction type is collinear not only with contrast but arguably also with topic and age constraints. Study 2, which introduces a dataset that is more richly annotated, was designed to tackle these concerns and to address RQ3 pertaining to the possible impact of the adolescent exodus away from Twitter around 2014.

Study 2. Investigating subject-*hun* in a smaller but richer dataset

Materials

As in Study 1, the dataset for Study 2 consists of materials sourced from the TwiNL-corpus, extracted with the same query as in Study 1, although this time we selected 250 token sets from five different timepoints: 2011, 2013, 2015, 2017, and 2019. For each timepoint, the subset represented a sample from the tweets containing *zij* or *hun* in that year. We restricted the search to: (1) tweets that were stored with geolocation; and (2) tweets whose senders' gender presentation could be determined with some confidence on the basis of their user profile or screen name. Most

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

importantly, the proportion between *zij* and *hun* in each sample was chosen to mirror the proportion between *zij* and *hun* in the population from which the sample was extracted (i.e., the total set of *zij*- and *hun*-tokens for each of the five timepoints). After removal of spurious hits, we ended up with 1,086 tokens, which manifested the *hun*/*zij*-distributions diagrammed in Figure 2.

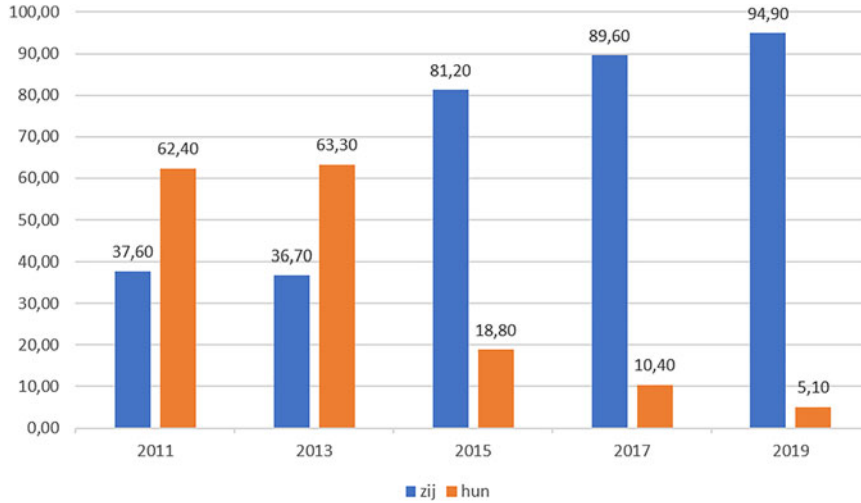

Figure 2. Relative frequency of standard *zij* and nonstandard *hun* as a function of year.

At first sight, Figure 2 seems to contradict the impression that subject-*hun* is diffusing in the Netherlands. If anything, the slight increase of the *hun*-proportion (to 63.3% in 2013) is dramatically reversed in 2015, with a proportion steadily decreasing toward 5.1% in 2019. Could this reversal reflect a change of demographics on Twitter, as suggested in Sanders (2023)?

Some diagnostics in the dataset of Study 2 seem to confirm a "changing of the guard" on Twitter. Table 4, which diagrams mean tweet length (in *n* of words) and the average number of spelling mistakes over the five time points, reveals that tweets doubled in size between 2011 and 2019 (going from 15.08 to 31.2 words,

Table 4. Mean tweet length (*n* of words) and mean *n* of errors as a function of year

	M Length	*M* Errors
2011	15.08	2.19
2013	14.61	2.16
2015	17.56	1.63
2017	19.67	1.36
2019	31.20	1.49

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

on average), and that they became prescriptively more compliant: both variables are arguably indicative of a progressive maturation of the Twitter community.

Even more suggestive are the data in Table 5, which quantifies the available hashtags in nine thematic categories: "Fictional media" comprises hashtags (like #GoT) that refer to television series like *Game of Thrones*; "Nonfictional media" refers to hashtags indexing talk shows and news programs (e.g., #dwdd for *De Wereld Draait Door*); "Politics" and "Sports" pertain to hashtags referring to political parties and sports teams, respectively; "Incidental groups" comprises hobby networks like #TeamToetjes 'Team Desserts'; "Evaluative" hashtags like #fail or #waanzin 'madness' are used to pass judgment on, for instance, Dutch railways and the COVID restrictions. The three final categories are "Culture" (#werelddierendag, 'Word Animal Day'), "Companies" (#Philips) and "Geographical locations" (#Amsterdam).

The steep increase in tweets about nonfictional television formats, political topics (often in combination with geographical locations), sports, and companies crucially reveals an authorship that is increasingly adult and educated from 2015 onward (recall that both of these demographics have been shown to shun subject-*hun*). Building on all this evidence, we can account for the reversal in *hun*-preference in Figure 2 in terms of what we will henceforth call the Great Exodus, namely, the migration of younger tweeters to other social media platforms.

Predictors and regression

Tokens were coded for four external and three internal predictors. External predictors included exodus, coded with values "pre" (2011, 2013) and "post" (2015, 2017, 2019). Gender was determined semiautomatically with a script that compared usernames with a list of ten thousand frequent first names (http://www.naamkunde.net/?page_id=293). The script, which was tested on a list of six hundred users whose gender could be uniquely determined, turned out to have a precision of 94.6% for full matches. Next, we manually checked each tweeter's gender and removed all tweets from the dataset from tweeters whose first name may have been ambiguous (as in

Table 5. Absolute frequency of nine hashtag categories as a function of year

	2011	2013	2015	2017	2019
Fictional media	3	0	0	2	0
Nonfictional media	1	3	4	7	18
Politics	1	4	5	8	10
Sports	2	1	6	9	9
Incidental groups	3	3	0	1	0
Evaluative	5	6	5	8	8
Culture	0	0	3	2	0
Companies	4	3	6	10	3
Geographical locations	0	1	7	7	9

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

the case of "Sam"). Gender was coded as a categorical distinction between M (male) and F (female). Length was coded ordinally with the value "short" for tweets between three and nine words long, "medium" for tweets between ten and nineteen words, and "long" for tweets of 20+ words; links, hashtags, and tags were excluded from this count. As an index of age, education, and formality, length is predicted to correlate negatively with a preference for *hun*.

As far as the internal *hun*-constraints are concerned, individuation was implemented as before, with the value "collect" for collective and "individ" for individuated reference (nineteen tokens in which *hun* or *zij* designated animals or nonanimate referents were removed from the dataset). Expressiveness was coded as before to distinguish between "zero," "one," and "twoplus" expressive strategies. Contrast, finally, was coded subjectively by three native speakers (the fourth author of this paper, one lay coder, and a third coder with a master's degree in Linguistics). A two-stage procedure was followed. In the first stage, the three coders classified tweets independently according to a three-level categorization "no/neutral contrast" (*they are not there*), "positive contrast" (*them are magnificent!*), and "negative" contrast (*them should bugger off!*); intercoder agreement at this stage was moderate (Fleiss's Kappa = .469). In the second stage, the first and fourth author collaboratively reclassified tweets on which no consensus was reached at the previous stage. Contrast was eventually entered as a categorical distinction between "Absent" for no or neutral contrast and "Present" for positive or negative contrast.

Next, we fitted a number of logistic mixed effect regressions on the data (*R*-package *lme4*, option *glmer* with the *logit* function; *R*-package *sjPlot*). The dependent variable modeled was the preference for *hun*. We considered all fixed effects and, in view of the very different distribution of *hun* before and after the exodus, all two-way interactions between gender, length, contrast, expressiveness, and individuation on the one hand and exodus on the other.

The best model contained significant main effects of all predictors, except construction type, which was not significant and did not improve model fit. None of the included interactions was significant or contributed to model fit. Figure 3 diagrams odds ratios and significance estimates for all predictors in the best model: recall that odds ratios on the right of the 1-mark index a positive effect on *hun*-use, while odds ratios on the left reveal a negative effect.

If anything, Figure 3 confirms Grondelaers and Van Hout's (2021) conclusion that most external predictors are much more important for model fit than internal predictors. As can be expected on the basis of the reversal in Figure 2, the effect of exodus has the highest impact: odds for *hun* are 1/0.10 = 10 times lower after 2013. In line with earlier findings (e.g., Grondelaers & Van Hout, 2021), we found a significant gender effect: odds for *hun* are 86% higher when a tweeter is feminine. As predicted, length also has a negative effect on *hun*-usage: odds for *hun* decrease 1/0.54 = 1.85 times in medium-length tweets, compared to short tweets, and they decrease 1/0.28 = 3.57 times in long tweets, compared to short tweets.

As far as the internal predictors of *hun* are concerned, the present analysis does not show any impact of construction type, our proxy for contrast in Study 1, which was a significant predictor there. Bivariate analysis *does* show an effect of construction type in Study 2 ($\chi^2 = 41.7$; $\varphi = 0.199$; $p = 0.000$), but this correlation

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

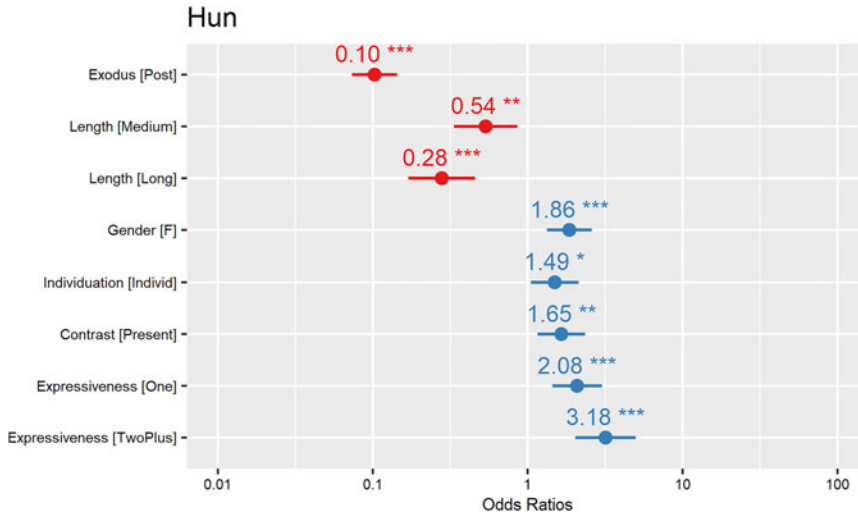

Figure 3. Model plot (with odds ratios and *p*-values) of logistic mixed effect regression on *hun*-preferences in the small dataset (* $p < .05$, ** $p < .01$, ***, $p < .001$).

between copular constructions and *hun* does not survive the competition with other predictors in the regression model. Plausible reasons for this are the decreasing proportion of copular constructions in the post-Exodus materials (which goes down from 25.3% in 2011 to 5.6% in 2019), and the availability of a better contrast predictor. The data in Figure 3 do confirm the crucial importance of dynamic self-stylization in terms of expressiveness: odds for *hun* increase by factor 2.08 as tweeters use one expressive strategy and by factor 3.18 as they use two or more expressive strategies. The effects of contrast (odds ratio = 1.65) and individuation (odds ratio = 1.48) are slightly smaller. The absence of any interaction of the main effects with exodus, finally, is a pivotal finding, for it demonstrates that the impact of our predictors is comparable in the pre- and post-Exodus era: the paucity of *hun*-tokens after 2014, consequently, does not seem to affect *hun*'s linguistic conditioning.

Discussion

In this section, we first interpret our findings in light of the methodological research questions put forward in the Research Questions and Hypotheses section, before we address the two hypotheses pertaining to the functioning of subject-*hun*.

The answer to RQ1 is a resounding yes: both studies show that grammatical vernacular choices abundantly recur in tweets (RQ1a). As a result, Twitter is an evident resource for (socio)linguists to obtain more balanced datasets featuring a sufficient number of nonstandard variants (RQ1b).

RQ2 requires a more nuanced answer. Yes, we can code for complex semantic predictors in the short format of tweets, but some empirical resourcefulness is needed. Coding for the semantic predictor contrast, for instance, necessitated ingenuity and prudence. In Study 1, we used a constructional proxy, the predicative copula template,

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

to delimit contrastive contexts, but this syntactic implementation is coarse-grained, and there is no perfect match between contrast and copular construction (RQ2a). As a result, we had to rely on the intersubjective contrast measure in Study 2, which is impractical to apply in very large datasets. RQ2b is even more problematic: the absence of independent access to a number of crucial *hun*-predictors can only partially be compensated for. Study 2 introduced a successful gender estimation algorithm, and Tables 2 and 3 in Study 1 have shown that we have limited, very indirect control over tweeter age: from specific hashtags and expressive strategies independently verified as adolescent resources, we were able to infer that copular constructions are preferred by comparatively younger tweeters. To date, no algorithms have been proposed to determine tweeters' education level.

RQ3 pertains to the fact that diachronic analysis on the basis of Twitter materials in The Netherlands may be compromised by the Great Exodus: Figure 2 demonstrated that the towering *hun*-proportions in the pre-Exodus period peter out after the Exodus. Although the decline of *hun* on Twitter is arguably a function of the demise of its preferred "biotope" (viz. young people's [aggressively] casual evaluations and exclusions of "other groups"), the main predictors of *hun* nevertheless remain alive and kicking after 2014, as revealed by the absence of significant interactions including exodus. Hence, the post-Exodus paucity of *hun*-tokens does *not* seem to affect the linguistic conditioning of subject-*hun* or, more generally, the suitability of Twitter materials for the investigation of this nonstandard variant.

We used Twitter data to test the hypotheses that *hun* is preferred because it is more suited (than its standard competitor) for contrastive evaluation (Hypothesis 1), and that this preference is sustained if not propelled by dynamic social meanings (Hypothesis 2). Both studies have confirmed the validity of these hypotheses. Our contrast implementation in Study 1 may have raised concern on account of its collinearity, but the independent implementation of contrast in Study 2 has allowed us to gauge the relation between contrast-profiling and dynamic self-stylization with more confidence: dynamic self-stylization appears to be the main determinant of *hun*-use (but read on for further elaboration).

A crucial methodological conclusion that can be drawn from both studies is that Twitter materials allow us to take the causality issue (i.e., to what extent is subject-*hun* not only *linked* to cool and dynamic social meanings but also *propelled* by it?) one step further. Recall that there is experimental evidence that subject-*hun* is associated with streetwise and dynamic meanings (Grondelaers et al., 2022), but the corpus evidence in the present paper demonstrates that social meaning plays a more decisive role than association: tweeters who self-stylize as cool dudes and gals are much more prone to a preference for *hun*-constructions. If anything, this finding demonstrates that social meaning (expression) is at the heart of variation and change processes, and that Twitter materials allow us to include such social meaning predictors in regression accounts of variation and change.

General Discussion

In this section, we first return to the interaction between construction type and expressiveness in Study 1 (following Figure 1), which crucially reveals that the *hun*-

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

boosting effect of the expressive self-stylization strategies is for the most part restricted to the noncopular constructions. Based on Tables 2 and 3, we had argued that the copula constructions in our dataset are plausibly preferred by younger tweeters, whereas noncopular constructions do not manifest this age restriction. If this interpretation is correct, the much more outspoken impact of expressive compensation strategies on the use of *hun* in the noncopular constructions signals that not young age, but youngish, informal self-stylization by tweeters of *any* age is the prime *hun*-booster. Noncopular cases in point are shown in (12)-(14).

12. ***hun*** *maken het laat, met veel overlast. mijn kinderen zijn erg vroeg en ze moeten hun energie toch ergens kwijt* 😊 *#buitenspelen #vroeg* 😊
'**them** stay out late, with a lot of nuisance. my children are very early (risers) and they have to get rid of their energy somewhere 😊 #playingoutside #early 😊'
13. *even een klacht want ze luisteren niet* 😊 *#anwb opgezegd maar* ***hun*** *zeggen van niet ! wij maken ook geen gebruik van hun !! #fail en betalen !!*
'A quick complaint because they are not listening 😊. #anwb [Dutch roadside assistance service] cancelled but **them** say I have not!! We do not use them either!! #fail and pay!!'
14. *verdiepen in geloof buurman ? ! ?* ***hun*** *moeten zich aanpassen, niet ik linkse tyfus hond. nep advocaat !! #wildersproces #pvv*
'acquire a deeper faith neighbor ? ! ? **them** have to adapt, not I you leftwing typhoid dog. fake lawyer!! #wilderstrial #pvv'

The content of examples (12)-(14) demonstrates that they are clearly produced by adults: (12) by a parent with noisy children; (13) by a car driver who has just cancelled his roadside assistance; and (14) by an indignant house owner/tenant who disagrees with his neighbor's leftwing political views. In all three cases, tweeters employ expressive strategies for humoristic or emotional stylization purposes: ironic understatement indexed by smileys and "dry" hashtags in (12), mildly agitated impatience indexed by the smiley, the excessive punctuation marking and the hashtags in (13), and the downright anger revealed by the hysterical punctuation marking ("? ! ?") and the swearwords "typhoid dog" and "fake lawyer" in (14). In none of these cases does *hun* signal young age, but in all cases, the use of *hun* crucially contributes to the stylistic effect.

A second point to be elaborated is the possibility that subject-*hun*'s semantic and social meanings are related. In Preston's (2019) account of the process that underlies the association of a variant and a specific social meaning, the persistent negative mediatization of *hun* as lowly educated and/or low-class would "imbue" the pronoun with these traits in the initial stages of its development. Since it is unlikely that such negative values would have supported *hun*'s diffusion across the Netherlands, we believe that the covertly prestigious exploitation of *hun* to flag indifference to social and linguistic norms (in the speech of cool soccer players and media figures) may have added a more positive "second-order indexicality" (in the terms of Silverstein, 2003) to the emerging iconic link. Crucially, the only difference between this extended social meaning and *hun*'s referential meaning is the "object" speakers distance themselves from: linguistic norms versus a group of other individuals.

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Conclusion

In this paper, we have reported two corpus analyses based on Twitter materials, collected to document the distribution and conditioning of nonstandard subject-*hun* in Netherlandic Dutch, but also to pressure-test the suitability of social media data for sociolinguistic analysis. We have come to three conclusions:

- Twitter is a prime resource to obtain large datasets of standard and nonstandard variants.
- Twitter datasets are much more diverse than the data typically used by sociolinguists, and the absence of independent demographic data necessitates caution and empirical resourcefulness. On that note, tweets are more suited to the analysis of variables constrained by language-internal predictors than phenomena conditioned by sociodemographic predictors.
- After having taken these issues into account (as much as possible), we found our Twitter corpora a suitable data source to settle a number of unresolved issues pertaining to the preference for *hun* as a subject. While the preference for *hun* is arguably motivated by the pronoun's contrastive and exclusionary charge, we have found that its preference is co-determined by its potential for dynamic and provocative self-stylization.

Our investigation inevitably suffers from a number of shortcomings. We did not exploit the full potential of gender and age assignment algorithms like *Python Gender Guesser*, and some dimensions of expressiveness can be tackled better through Sentiment Analysis (i.e., the use of computational tools to automatically extract and quantify affective states and subjective information, which mainly originated in the marketing domain but has been applied in (socio)linguistics [e.g., Aboada, 2016]).

An unavoidable consequence of the sociolinguistic focus on both internal, external, and prestige-related predictors, is the necessity of more hand-coding, and the concomitant need to reduce the size of the dataset, which, in turn, challenges the main advantage of Twitter materials, that is, the almost unlimited availability of large datasets featuring nonstandard variants. In this sense, the reliance on tweets for sociolinguistic analysis will always remain a trade-off between size and detail. However, we hope to have shown that even when one focuses on size, Twitter data can go a long way toward resolving micro-sociolinguistic questions.

Acknowledgments. We are indebted to Mirthe Koppenberg for her invaluable assistance in the coding of (part of) the corpus, and to Astraea Blonk for writing the gender estimation script in Study 2.

Competing interests. The authors declare none.

Note

1. These verb forms are *denken* 'think'; *doen* 'do'; *gaan* 'go'; *hadden* 'had'; *hebben* 'have'; *komen* 'come'; *krijgen* 'get'; *kunnen* 'can'; *liggen* 'lie'; *maken* 'make'; *moeten* 'must/have to'; *mogen* 'may/can'; *staan* 'stand'; *vechten* 'fight'; *vinden* 'find'; *waren* 'were'; *weten* 'know'; *willen* 'want'; *worden* 'become'; *zeggen* 'say'; *zijn* 'are'; *zitten* 'sit'; *zullen* 'shall/will.'

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

References

Aboada, Maite. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2:325–347.

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In T. Kristiansen & N. Coupland (eds.), *Standard languages and language standards in a changing Europe*. Oslo: Novus. 145–161.

Bohmann, Axel. (2016). Language change because Twitter? Factors motivating innovative uses of because across the English-speaking Twittersphere. In L. Squires (ed.), *English in computer-mediated communication. Variation, representation, and change*. Berlin: De Gruyter Mouton. 149–178.

Brody, Samuel, & Diakopoulos, Nicholas. (2011). Cooooooooooooooollllllllllll!!!!!!! Using word lengthening to detect sentiment in microblogs. *Proceedings of the* 2011 *Conference on Empirical Methods in Natural Language Processing*, 562–570.

Brown, Earl K. (2016). On the utility of combining production data and perceptual data to investigate regional linguistic variation: The case of Spanish experiential *gustar* 'to like, to please' on Twitter and in an online survey. *Journal of Linguistic Geography* 3:47–59.

Carey, Stan. (2013). "Because" has become a preposition, because grammar. *Sentence First*. http://stancarey.wordpress.com/2013/11/13/because-has-become-a-preposition-becausegrammar/ (accessed March 14, 2016).

Coats, Steven. (2016). Grammatical feature frequencies of English on Twitter in Finland. In L. Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Berlin: De Gruyter. 179–210.

De Decker, Benny, & Vandekerckhove, Reinhild. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51:253–281.

Foulkes, Paul, & Docherty, Gerard. (1999). *Urban voices: Accent studies in the British Isles*. London: Arnold.

Garber, Megan. (2013). English has a new preposition, because Internet. *The Atlantic*. http://www.theatlantic.com/technology/archive/2013/11/english-has-a-new-prepositionbecause-internet/281601/ (accessed March 14, 2016).

Grondelaers, Stefan, & Marzo, Stefania. (2023). Why does the *shtyle* spread? Street prestige boosts the diffusion of urban vernacular features. *Language in Society* 52:295–320.

Grondelaers, Stefan, Van Gent, Paul, & Van Hout, Roeland. (2022). On the inevitability of social meaning and ideology in accounts of syntactic change. Evidence from pronoun competition in Netherlandic Dutch. In T. Karoli Christensen & T. J. Jensen (eds.), *Explanations in sociosyntax. Dialogue across paradigms*. Amsterdam: Benjamins. 120–143.

Grondelaers, Stefan, & Van Hout, Roeland. (2021). From Big Brother to IKEA. The quest for the optimum data source to study a notorious pronoun diffusion in Dutch. In G. Kristiansen, K. Franco, S. De Pascale, L. Rosseel & W. Zhang (eds.), *Cognitive sociolinguistics revisited*. Berlin: De Gruyter Mouton. 371–383.

Haddican, Bill, & Johnson, Daniel Ezra. (2012). Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18:31–40.

Hilte, Lisa, Vandekerckhove, Reinhild, & Daelemans, Walter. (2018). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23:293–323.

Jones, Taylor. (2015). Toward a description of African American Vernacular English dialect regions using "Black Twitter." *American Speech* 90:403–440.

Oostdijk, Nelleke. (2003). Het Corpus Gesproken Nederlands: Veelzijdig onderzoeksinstrument voor o.a. taalkundig en taal- en spraaktechnologisch onderzoek. *Link* 14:3–6.

Ordelman, Roeland, De Jong, Franciska, Van Hessen, Arjan, & Hondorp, Hendri. (2007). TwNC: A multifaceted Dutch news corpus. *ELRA Newsletter* 12:4–7.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5:46–54.

Preston, Dennis R. (2019). How to trick respondents into revealing implicit attitudes – talk to them. In L. Rosseel & S. Grondelaers (eds.), *Implicitness and experimental methods in language variation research*. Special issue of *Linguistics Vanguard* 5(1).

Rehbein, Ines. (2014). Using Twitter for linguistic purposes: Three case studies. Paper presented in the workshop *Web Data as a Challenge for Theoretical Linguistics and Corpus Design* of the 36*th Annual Conference of the German Linguistic Society*, March 2014.

https://doi.org/10.1017/S0954394523000121 Published online by Cambridge University Press

Sanders, Erik. (2023). *Vox Populi. On forecasting elections with Twitter*. Doctoral dissertation, Radboud University Nijmegen.

Silverstein, Michael. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication* 23:193–229.

Sneller, Betsy, & Roberts, Gareth. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition* 170:298–311.

Stuart-Smith, Jane, Gwilym Pryce, Claire Timmins, & Barrie Gunter. (2013). Television can also be a factor in language change: Evidence from an urban dialect. *Language* 89:501–536.

Suciu, Peter. (2022). Millennials threaten to quit Musk-owned Twitter. Retrieved from https://www.forbes.com/sites/petersuciu/2022/11/17/millennials-threaten-to-quit-musk-owned-Twitter/ on February 28, 2023.

Tjong Kim Sang, Erik, & Van den Bosch, Antal. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal* 3:121–134.

Turpijn, Loes, Kneefel, Samantha, & Van der Veer, Neil. (2015). *Nationale social media onderzoek 2015*. Amsterdam: Newcom Research & Consultancy.

Van Bergen, Geertje, Stoop, Wessel, Vogels, Jorrig, & De Hoop, Helen. (2011). Leve Hun! Waarom hun nog steeds hun zeggen. *Nederlandse Taalkunde* 16:2–29.

Van Hout, Roeland. (2003). Hun zijn jongens. Ontstaan en verspreiding van het onderwerp 'hun.' In J. Stroop (ed.), *Waar gaat het Nederlands naartoe? Panorama van een taal*. Amsterdam: Bert Bakker. 277–286.

Verheijen, Lieke. (2018). Orthographic principles in computer-mediated communication: The SUPER-functions of textisms and their interaction with age and medium. *Written Language and Literacy* 21:111–145.

Vor der Hake, Jan A. (1911). Is de beleefdheidsvorm U 'n verbastering van UEd.? *De Nieuwe Taalgids* 5:16–34.

Cite this article: Grondelaers S, van Hout R, van Halteren H, Veerbeek E (2023). Why do we say *them* when we know it should be *they*? Twitter as a resource for investigating nonstandard syntactic variation in The Netherlands. *Language Variation and Change* **35**, 223–245. https://doi.org/10.1017/S0954394523000121