Luca Bevacqua* and Tatjana Scheffler

Form variation of pronominal it-clefts in written English

A corpus study in Twitter and iWeb

https://doi.org/10.1515/lingvan-2019-0066

Abstract: Clefts are well-studied as a construction which induces emphasis on its clefted referent. However, little is known about the distribution of different stylistic forms of it-cleft variants. We report on a corpus study mining data from Twitter, targeting sentences clefting a pronoun in English. We examine the following features: case and syntactic role of the clefted pronoun, contraction of the copula, choice of complementiser and use of emphasis markers. The results show systematic associations between these features. A further comparison between the Twitter dataset and data from iWeb, a corpus of general-use web language, shows significant differences in levels of emphasis and formality, positioning Twitter language in the middle of the conceptual orality spectrum.

Keywords: clefts, corpus study, English, emphasis, computer-mediated communication

1 Introduction

Cleft constructions such as (1–3) are used to focus the clefted referent, enhancing its prominence and making it more accessible for re-mention in subsequent discourse.

(1) It's she that this post honors.[1]

(2) It is him who has caused her to smile brightly.

(3) What John lost was his keys. (Prince, 1978)

There are well-studied and clear pragmatic differences between English it-clefts such as (1–2), and wh-clefts (3) (Prince 1978). In contrast, there is less existing research discussing possible differences between form variants: the use of contraction, case of the pronoun and complementiser vary between sentences (1) and (2). In this paper, we investigate these form variants in more detail. In particular, we want to determine which means are used to mark or emphasise the clefted referent. To do this, we study naturally occurring it-clefts in data extracted from the social media platform Twitter: this allows us to tap into relatively spontaneous utterances containing clefts as a focussing construction. We show that the observed variation is by no means random: instead, we identify formality as a driving factor for expressing emphasis in the Twitter clefts, which leads to three preferred variants of it-clefts in English spontaneous written text. We compare our findings with another

1 All examples are attested sentences from Twitter, unless marked otherwise (see Section 3.1 for details on the data collection). We quote only the cleft sentence, using the original spelling and punctuation. Twitter does not allow the distribution of aggregated tweet corpora, but our data set containing the extracted it-clefts, along with the original tweet IDs and our label annotations, is available in the supplementary materials, along with the R code we used for the analysis.

***Corresponding author: Luca Bevacqua**, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK, E-mail: lbevacqu@ed.ac.uk
Tatjana Scheffler, German Department, Ruhr-Universität Bochum, Bochum, Germany; and Department of Linguistics, University of Potsdam, Potsdam, Germany, E-mail: tatjana.scheffler@rub.de. https://orcid.org/0000-0001-7498-6202

corpus of web data, i.e. iWeb (Davies 2018), containing more strongly edited but still variable content. This corpus shows much less variation, because standards promoted in editing guidelines result in an almost categorical preference of certain forms. Our systematic investigation of a relatively rare phenomenon in a big dataset opens up new directions for explaining expressions of focus, emphasis, and the formality gradient in English.

In the following section, we report some findings on it-clefts, emphasis, and formality from the literature (Section 2). Afterwards, we present our corpus study, first within Twitter, and then as a comparison between Twitter and iWeb (Section 3). We note a systematic association of more informal (contraction, accusative-case pronouns, *that* as a complementiser and emphasis markers) and formal (uncontracted clefts, nominative case, and *who*) features, as well as a general wider variation of these features in Twitter writing than in iWeb's, positioning Twitter in a more central position in the "conceptual orality" spectrum (Koch and Oesterreicher 1985). Finally, we discuss these results in the light of the background literature and rules of prescriptive grammars (Section 4 and 5).

2 Background

2.1 It-clefts

If a focus is pragmatically needed, e.g. to contrast a piece of information, it will usually be grammatically marked, either prosodically or morphosyntactically, or both (Lambrecht 1994: p. 64). The expression of this marking can eventually be conventionalised, and cleft constructions are a good example of such grammaticalisation.

English it-clefts consist of the pronoun *it*, followed by a form of the verb *to be*, a clefted constituent, and a complementiser, which introduces a relative clause that is attributed of the clefted phrase (Lambrecht 2001). It-clefts introduce two meaning parts: (i) a presupposition that the property in the clause following the complementiser holds of some entity; and (ii) an assertion (at-issue contribution) that this property holds of the entity denoted by the clefted constituent. We will thus call the clause following the complementiser the *presupposed clause*, following Prince (1978); the clefted constituent is also called the focussed constituent.

The literature on clefts has concentrated on several central questions of syntax and semantics. On the one hand, research has addressed the issue of the underlying syntactic structure of cleft sentences. On the other hand, given that the truth conditions between a cleft sentence and its corresponding simple clause do not differ, there has been discussion about the exact pragmatic contribution of using a cleft construction, especially with regard to information structure (Delin 1992).

Here, we focus on the form of the cleft construction itself, which offers a lot of variability within the frame defined above. We identify the following "moving parts" of it-clefts:

(4) a. **Contraction:** The copula *be* may appear contracted or uncontracted.
b. **Case:** The clefted constituent may be in nominative or accusative case. In English, this feature is only visible in pronouns.
c. **Complementiser:** The complementiser in the relative clause can be *that* or a *wh*-complementiser. In the case of clefted objects, the complementiser may even be empty (Quirk et al. 1985: p. 1385), but we did not consider this particular case.[2]

The example in (1) illustrates a contracted, nominative cleft with the complementiser *that*, whereas (2) is an uncontracted, accusative cleft with *who* as the complementiser.

Of these features of variation in clefts, Maier (2013, 2014) studies the case of the pronoun using corpora. Maier (2013) shows that the form of the pronoun is influenced not only by its position in the

2 Object clefts are rare, see discussion below in Section 3.3.

sentence, but also by its function, with variation related to the discourse medium: while written language prefers nominative-case pronouns, spoken English uses more accusatives. Maier (2014) accounts for the variation in case with information structural rationale, claiming that the nominative case of English pronouns in clefts was refunctionalised to signal emphasis (which he calls "focus"). We note that since he investigates both it-clefts and other related constructions together, the conclusion that it-clefts with nominal case pronouns are more emphatic than the accusative ones does not strictly follow from his statistical analysis.

The case of the pronoun in it-clefts is also mentioned by Patten (2010, 2012). Patten notes how the case of the pronoun is fundamental to the type of syntactic account: under an expletive account (in which the cleft is used to mark a sentence with the addition of a "dummy" element), the case of the postcopular element should be nominative, but this pattern "is not found in the dialect of the majority of speakers" (Patten 2010: p. 146, reporting a finding in Akmajian 1970). However, using the ICE-GB corpus (Nelson et al. 1998), Patten (2012) shows that clefted pronouns in the nominative are almost twice as frequent as those in the accusative (with a proviso of very low counts for either), and agrees with Akmajian (1970) in interpreting this as the preference for a prestige form, which gives an appearance of formality. This is further confirmed by a corpus study in Biber et al. (1999), where clefts with a nominative pronoun and *who* are frequent in fiction and news despite a general predominance of the accusative form. This is motivated claiming that nominative forms may be perceived to be more correct, possibly because of prescriptive rules moulded on those of Latin grammar.

As for complementisers, Collins (1991) finds that about two-thirds of it-clefts contain *that*, but most other *wh*-words are viable for the role. Cheshire et al. (2013) conduct an in-depth investigation of complementisers in restrictive relative clauses. They find that *who* is used almost exclusively in subject relative clauses. The choice of complementiser shows a strong variation in their data, though for younger speakers, most instances are realised with *that*. The complementiser *that* is the most general (lacking the animacy requirement of *who*). For some speakers, Cheshire et al. (2013) conclude that the subject preference of *who* is being grammaticalised as (additional) "topic" marking, in the sense that referents relativised with *who* are more salient (therefore, more available for re-mention or more "topic-worthy") in subsequent discourse.

Looking at their distribution across registers, Collins (1991) notes that clefts outnumber pseudo-clefts in writing, while the opposite happens in speaking. Within spoken genres, it-clefts are more used in prepared monologues than in dialogues, further pointing in a direction of higher formality. However, internal features such as the case of the pronoun or the complementiser of the cleft are not analysed.

We are not aware of studies that systematically investigate multiple axes of variability in English it-clefts. In this study, we will consider these variants within a theory of emphasis marking.

2.2 Emphasis

We define emphasis as a linguistic means chosen by a speaker to direct a hearer's attention, for example to help single out an intended referent (Zimmermann 2008). This concept has been inconsistently referred to throughout the pragmatics literature, being at times likened to prosodic stress (e.g. Heath 2018) and often used interchangeably with other under-defined or theory-specific terms such as "prominence" or "salience". Clefts of all kinds are emphatic in the sense that, as reflected by their prosodic prominence (Lambrecht 2000), they identify the clefted element as something that the speaker considers especially worthy of attention, because of its status as new or contrastive information. In addition to the grammaticalised emphatic function of clefts, syntactic or morphological (or indeed prosodic) cues exist and can be used by the speaker to indicate particular emphasis.

Computer-mediated communication, being written, does not mark emphasis prosodically. However, it makes available a large range of other emphasis markers, that have been shown to take information structural roles. McAteer (1992) studies how the alteration of typeface can mark information focus in a way completely akin to intonational signals. Some of these are standard in writing and have conventionalised functions, such

as punctuation or paragraphing, others are not, like all caps or boldface. These functions are included by McAteer (1992), along with clefting, in a set of "foregrounding strategies" that a writer can employ to mark contrast and emphasis. In more recent work on writing in social media, Heath (2018) identifies non-standard capitalisation as a marker of focus and emphasis, while Scott (2015) analyses hashtags as topic markers.

To study the variable encoding of the cleft construction and its relation to emphasis marking, we annotate these types of emphasis phenomena in our data:

(5) Types of emphasis marking:
Capital pronoun: The clefted pronoun is capitalised: e.g. "it's HE that's threatened us"
Capital initial: The first letter of the clefted pronoun is capitalised: e.g. "it is She who has to live with the consequences"
Capital cleft: The whole cleft is capitalised: e.g. "it's HIM WHO IS THE PROBLEM"
Exclamation: The cleft contains exclamation points: e.g. "it is he who holds all the baggage!"
Non-standard comma: Non-standardly placed comma to mimic prosody: e.g. "it is he, that is my one true love"
Asterisks: Asterisks around the clefted pronoun: e.g. "it's *he* who did it"
Other: Other strategies, such as tildes or inverted commas around the clefted pronoun.

Social media has been shown to innovate written language systematically, with a big influence coming from spoken language: for example, spelling on Twitter appears to follow phonological variations relatively frequently (Eisenstein 2013a). Non-standard emphasis markers are, likewise, more widespread in social media than in standard written language.

2.3 Formality

Another dimension which interacts with emphasis, in particular in social media, is formality. Heylighen and Dewaele (1999) define expressivity of language (which is another reshaping of emphasis in that it corresponds to linguistic genres with high detail and emotionality) as well as formality in terms of the orthogonal variability axes fuzziness and context-dependence. In this view, context-dependence refers to the amount of inference that has to be made in order to interpret an expression, fuzziness refers to the opposite of semantic precision of expression. Formal style will tend to minimise fuzziness and context-dependence, while expressive language is precise (not fuzzy) but highly context-dependent.

In this account, expressivity is a linear combination of precision (-fuzziness) and context-dependence: for example, a proper name will be much less fuzzy than a pronoun, which will in turn be highly context-dependent. Likewise, the two complementisers will vary along the fuzziness axis: *who*, with its [+animate] trait, includes more information than *that*, and being more precise it will be more formal and more expressive. Its use could thus correlate with the avoidance of contraction in the cleft sentence (which increases its phonological weight and thus the formality of the register: see Kjellmer (1997) for extensive examples). Similarly, the use of emphasis markers indicates increased expressiveness and could thus be associated with uncontracted clefts, or with a more precise form of the complementiser like *who*.

Further, it is known that social media are a highly informal medium, often reflecting features frequently found in spoken conversations more than written texts (cf. Tagliamonte and Denis 2008; Storrer 2013; Scheffler 2017). This places social media texts in the middle of the "conceptual orality" spectrum (Koch and Oesterreicher 1985) which moves along an axis of medium (spoken to written) and one of formality (informal to formal). Since emphasis by itself is also an effect of markedness, the specific use of a formal variant in social media would constitute a marked case and thus increase the overall emphasis of the utterance. In the following, we investigate the use of formal and emphatic variants of it-clefts in Twitter data.

3 Empirical study

We present a systematic empirical study which aims to classify it-cleft variants along emphasis and formality dimensions. It-clefts are a rare phenomenon in spoken and written corpora: for example, the iWeb corpus (Davies 2018) only contains about 4000 of them, a very low number for a corpus including texts from more than 22 million web pages. In addition, a well-founded comparison of naturally occurring it-clefts from corpora should aim to exclude factors which can potentially confound the information structure, syntactic variation, and interpretation of the utterance. In particular, this concerns the syntactic and prosodic type of the clefted phrase, as well as its meaning (for example, length and weight of the phrase, animacy of the referent, etc.). We therefore choose to look only at it-clefts with clefted third-person pronouns. Unfortunately, this restriction makes most existing corpora too small for statistical analysis. Mining Twitter for data, conversely, still provides a good amount of examples to study. In addition to data availability, Twitter as a medium has another advantage for studying this phenomenon: as indicated above, since "standard" language norms are less applicable here, we expect increased variability in the observed forms of structures such as clefts.

We first mined tweets with it-clefts and one of the authors annotated them (see Section 3.1). After an initial analysis, new Twitter data was collected to internally validate the findings using the same analysis on a second sample. Finally, we extracted clefts from the iWeb corpus (Davies 2018), replicated the analysis in its data, and compared the results with the Twitter data.[3]

3.1 First data collection and annotation

Tweets including at least one cleft were collected on 25th to 29th March 2019 using TAGS (Hawksey 2016). The queries were url-encoded to target specific tweets without excluding other factors such as emphasis markers or punctuation, and excluded retweets.[4] The features targeted in the searches are listed in (4): contraction, case of the pronoun, and complementiser. This means all it-clefts introduced with or without a contraction (*it's*[5] and *it is*), with a third person singular pronoun in either nominative or accusative case, and *that* or *who* as a complementiser, followed by a presupposed clause of any length.

All tweets using a cleft in the context of a quote from sacred scriptures (e.g. "But thou shalt remember the LORD thy God: for it is he that giveth thee power to get wealth") were manually removed from the dataset because they are representative of a different register of language. Many of these appeared multiple times as "quotes of the day". This left the dataset with a total of 798 tweets that conform to our constraints.[6]

The data was manually annotated for the following features:

Number of referents: 1 for clefts wherein only one referent was mentioned (e.g. "it's he that needs to go"), 2 for clefts involving two referents (e.g. "it's she that called him").

Role of the clefted element: Subject (e.g. "it is he that assisted Castro") versus Object (e.g. "it is he that I will respond to"). This annotation only concerns clefts with two referents (i.e. clefts with both a subject and an object).

Emphasis: presence of emphasis markers: see Section 2.2.

Cleft: text of the cleft(s), isolated from the rest of the tweet.

45 tweets (5.6% of the total) contained multiple clefts. Only the first cleft of every tweet was considered in the following analysis.

---

3 We also obtained the English portion of the large multilingual spoken cleft corpus extracted from European Parliament recordings (Bouma et al. 2010). However, the number of clefts that match our criteria was too small to replicate our analysis directly.

4 All queries are reported in Appendix I.

5 We thank an anonymous reviewer for pointing out that the informal spelling variant its, lacking the apostrophe, is not matched by our queries, thus excluding this group of (plausibly) informal tweets from the data set. While this does potentially skew the overall view of our Twitter data towards increased formality, we chose not to include this variant as it is virtually absent from edited writing like that in the iWeb corpus, making the comparison of the two genres not possible.

6 The raw numbers for each feature are reported in Appendix II. The full data set can be consulted in the supplementary materials to this article.

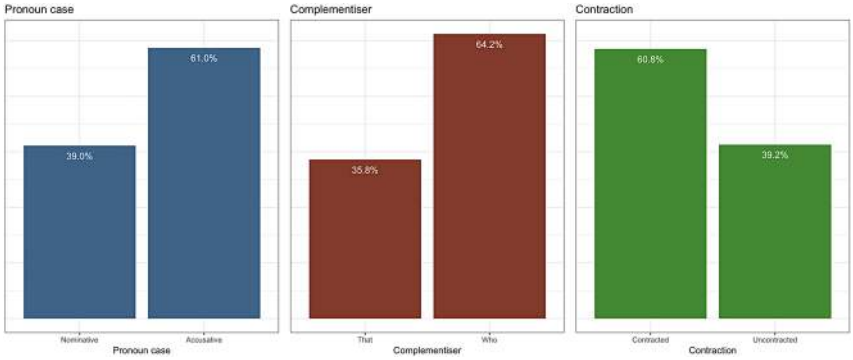

Figure 1: Ratios of the three features in the data set.

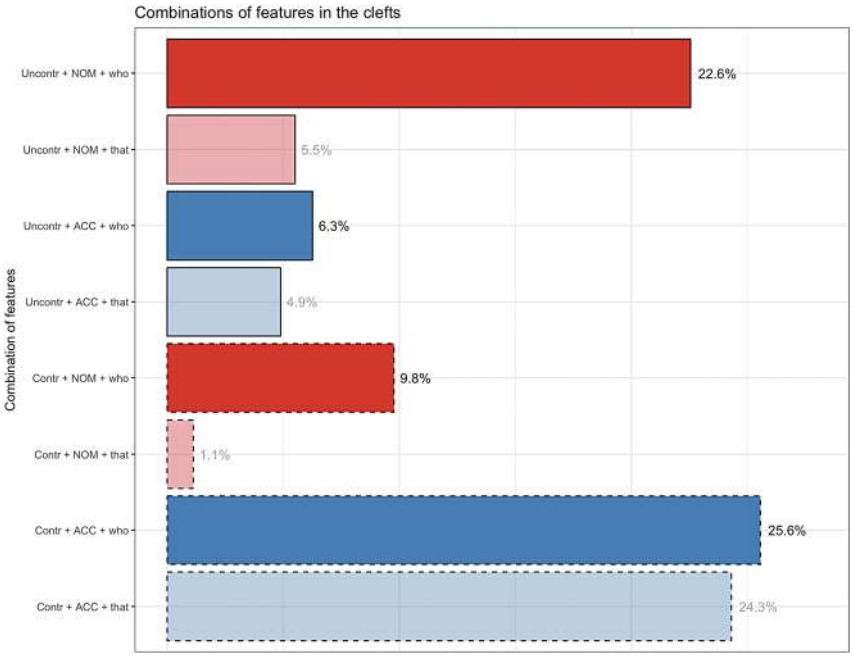

Figure 2: Distribution of the eight feature combinations in the data. The outline of the bars distinguishes uncontracted (continuous line) from contracted (segmented line) clefts, the color of the bar distinguishes the case of the pronoun (nominative = red, accusative = blue), and the intensity of the color distinguishes the complementiser (*who* = full color, *that* = transparency).

3.2 Twitter data analysis

The first analysis targets the full dataset and some of the features annotated in the tweets. The basic distribution of the three main factors listed in (4) (contraction, pronoun case and complementiser) is shown in Figure 1. It can be gathered that in general clefts in English-language tweets are more often contracted and contain an accusative pronoun with the complementiser *who*. Moreover, more clefts with masculine than with feminine pronouns are produced (489 vs 309, 61.3 vs 38.7%; this ratio is not visualised).

Figure 2 shows the distribution of all eight combinations of the main features considered. It seems evident that the features are not distributed randomly: three combinations dominate the picture. About equally frequent, they together account for about 3/4 of the it-cleft occurrences. These are exemplified in (6):[7]

(6) **Uncontr-NOM-who:** it is she who has the power
Contr-ACC-who: it's her who has to do it
Contr-ACC-that: it's him that is the target

7 Note that all pronouns are syntactic subjects.

In the following, we statistically evaluate the correlations between the individual features. Given that no direction of effect can be discerned, chi-squared tests[8] were used to test for the associations and Cramér's V for their strength. Cohen's (1988) guidelines were used to evaluate the strength of the effects, with $0.1 < V < 0.3$ indicating a small, $0.3 < V < 0.5$ a medium, and $V > 0.5$ a large effect size. We report only significant effects, except where noted otherwise.

The **case of the pronoun** is associated with both the use of contraction ($\chi^2 = 230.02$, $p < 0.0001$, $V = 0.54$: large effect size) and the presence of emphasis markers ($\chi^2 = 12.47$, $p < 0.001$, $V = 0.13$: small effect size). This means that clefts are more contracted with accusative pronouns and less with nominative pronouns, and that emphasis markers co-occur significantly more often with nominative case pronouns. These effects are shown in Figure 3.

The **choice of complementiser** is associated with the case of the pronoun: nominative pronouns are used more frequently with *who* ($\chi^2 = 78.31$, $p < 0.0001$, $V = 0.31$: medium effect size). Moreover, the complementiser *who* co-occurs significantly more with uncontracted clefts ($\chi^2 = 19.46$, $p < 0.0001$, $V = 0.16$: small effect size). The two effects are visualised in Figure 4.

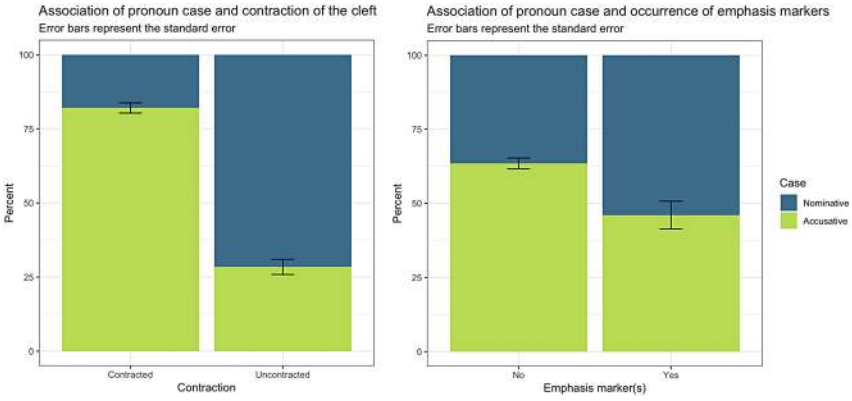

Figure 3: Associations of pronoun case with contraction and emphasis. Figures 3–6 show the co-occurrence of two features, represented on the *x* axis and through colour. The *y* axis shows the proportion of each combination.

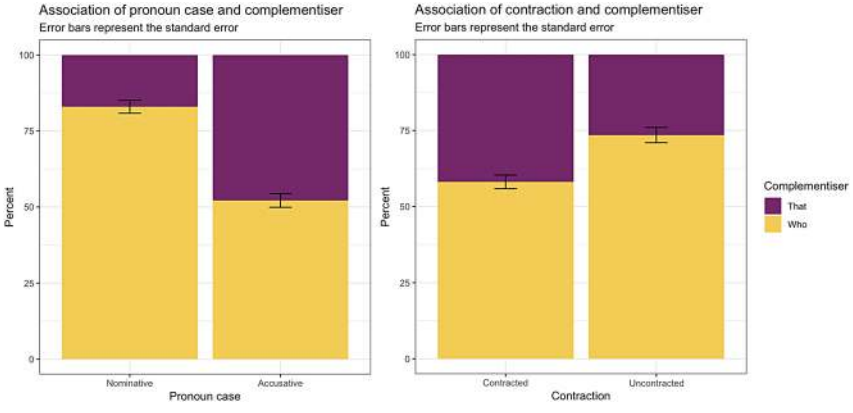

Figure 4: Associations of complementisers with case and contraction.

8 All *df* = 1, except when a Monte Carlo simulation is used (in this case, *df* are not applicable).

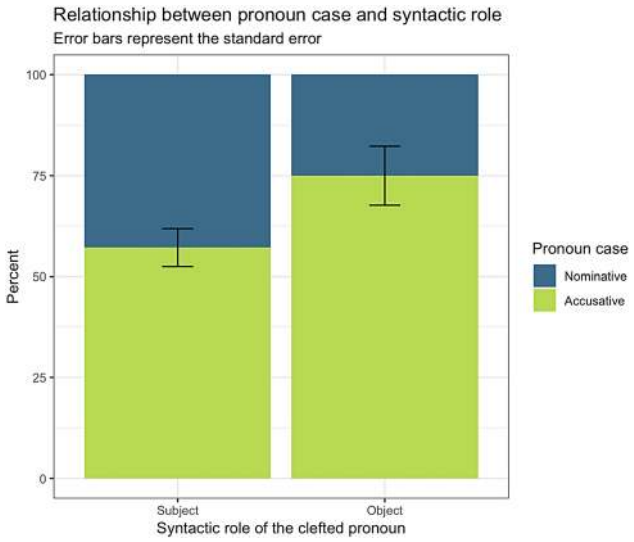

Figure 5: Distribution of case and syntactic role in clefts with two referents. The difference between subjects and objects is not statistically significant.

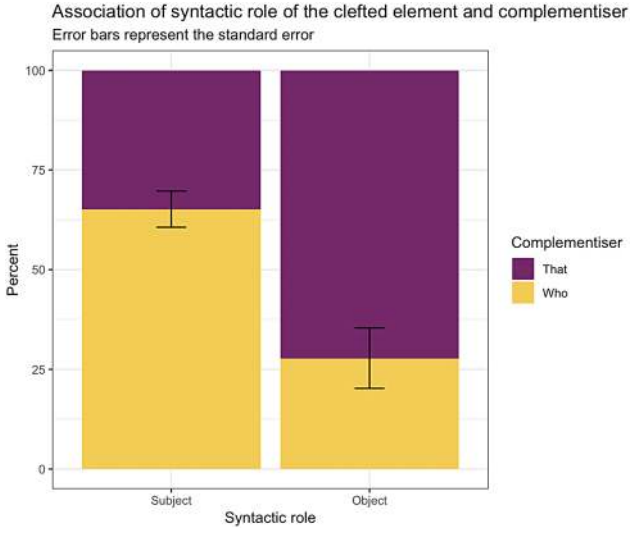

Figure 6: Associations of complementisers with the syntactic role of the clefted element.

3.3 Syntactic role of clefted pronouns

Subsetting the data to the 148 clefts involving two referents, an obvious prevalence of clefts of the subject can be seen (112 vs 36, 75.7 vs 24.3%). While the subject is naturally a topical element in the information structure of an utterance, the object often provides new information, making it less necessary to mark explicitly with a focus.

In structural positions, the pronoun case is determined by the syntactic role the pronoun occupies (with nominative for subjects, and accusative for objects). However, this is not found in the clefted positions in our data. There is no significant association of the syntactic role of the clefted element with the case of the pronoun ($\chi^2 = 3.67$, $p = 0.06$). As shown in Figure 5, both subjects and objects are more likely to appear in accusative case in our Twitter data.

No significant association of syntactic role with the use of contraction in the cleft was found ($\chi^2 = 0.004$, $p = 0.95$), nor with the presence of emphasis markers ($\chi^2 = 1.5$, $p = 0.22$). On the other hand, the syntactic role

Table 1: Ratios within the target features in the two batches of data.

	First batch	Second batch
Contracted | Uncontracted	60.8% | 39.2%	65.7% | 34.3%
Accusative | Nominative	61.0% | 39.0%	60.7% | 39.3%
That | Who	35.8% | 64.2%	42.6% | 58.4%

Table 2: Comparison of χ^2 and Cramér's V tests on the two batches of data. All $p < 0.001$.

	First batch	Second batch
Contraction ~ Case	$\chi^2 = 230.02$, $V = 0.54$	$\chi^2 = 261.99$, $V = 0.65$
Complementiser ~ Case	$\chi^2 = 78.31$, $V = 0.31$	$\chi^2 = 89.68$, $V = 0.38$
Contraction ~ Complementiser	$\chi^2 = 19.46$, $V = 0.16$	$\chi^2 = 27.95$, $V = 0.21$

is associated with the complementiser: clefts of the subject co-occur significantly more with *who* rather than *that* ($\chi^2 = 15.47$, $p < 0.0001$, $V = 0.32$: medium effect size). This is visualised in Figure 6.

3.4 Replication

The associations above were confirmed with a second, not manually annotated batch of 616 tweets mined on eighth May 2019. Table 1 compares percentages in the two datasets. Three of the chi-squared and Cramér's V tests could be replicated on the non-annotated data, yielding similar results (see Table 2). This replication shows that the results from our first analysis are stable across data samples.

3.5 Comparison between Twitter and general web English

These associations were then recalculated on the results of corpus searches in the iWeb Corpus (Davies 2018), a 14-billion-word corpus of written English data from 94k+ systematically-chosen websites.

Within the 4056 pronominal it-clefts found in the iWeb corpus, all associations seen in the Twitter data are replicated (with $p < 0.001$). Case is associated with contraction ($\chi^2 = 783.24$, $V = 0.44$) and with complementiser choice ($\chi^2 = 362.28$, $V = 0.3$), and contraction is associated with complementiser choice ($\chi^2 = 89.15$, $V = 0.15$). While the effects are more significant, likely because of the dimensions of the corpus, Cramér's V figures show slightly smaller effects (especially for the association of contraction and pronoun case). Note that same direction of the effects will be confirmed through logistic regression.

The Twitter and iWeb results were compared through chi-squared tests in a 2×16 contingency table (including gender among the variations), computing p-values by Monte Carlo simulation (Hope 1968) because of the low counts for some of the structures. A significant difference was found ($\chi^2 = 1827.5$, $p < 0.001$) between how Twitter and general-purpose (web) English use clefts of different types. Figure 7 shows how Twitter users adopt more accusative pronouns and contractions, while all-purpose language in iWeb attests at a more formal level avoiding contraction, as dictated by some style manuals, e.g. the Wikipedia Manual of Style (Wikipedia contributors 2019). Nominative pronouns are used more outside of Twitter, mostly because of the very high frequency of the structure *It is he who* (forming 77% of all nominative-pronoun structures). Complementisers are used similarly across genres.

A generalised mixed-effects logistic regression (Bates et al. 2015) was fitted to predict the "genre" binary (Twitter vs iWeb) from the features of the clefts, i.e. case, contraction and complementiser, with a random effect for the gender of the pronoun (not assumed to display systematic variation across genres). All predictors were

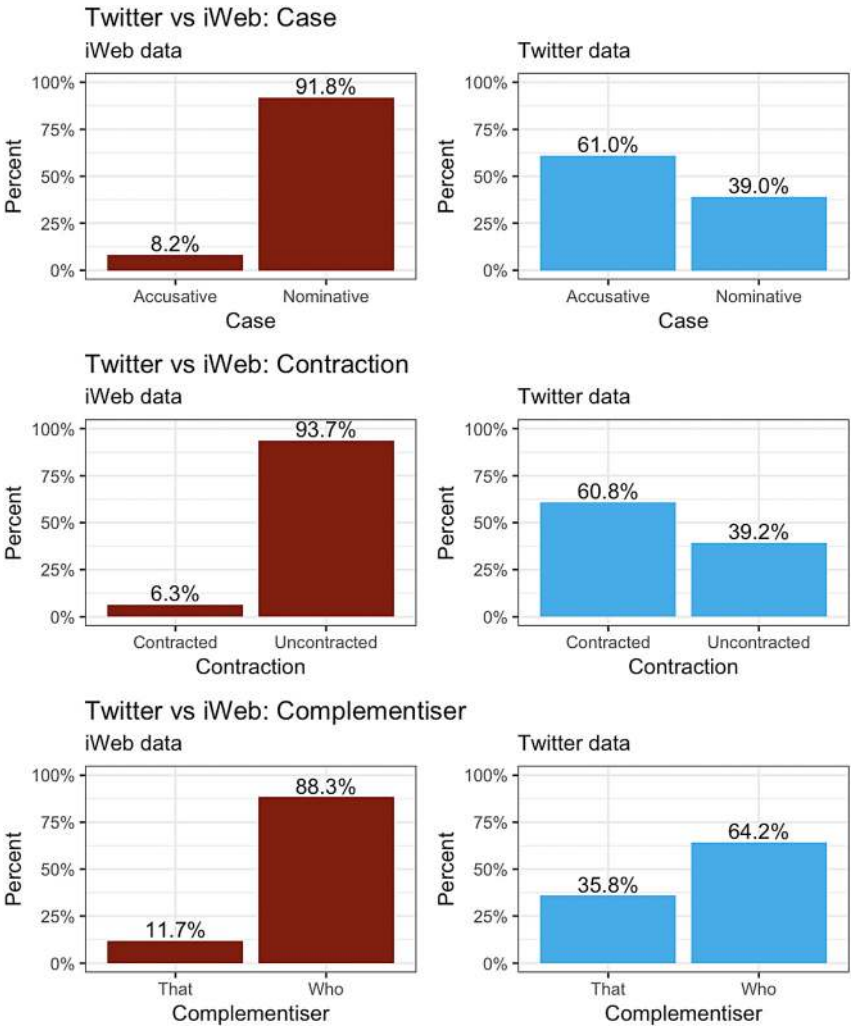

Figure 7: Comparison of the iWeb corpus with Twitter. The features on the right-hand side of each plot are those more heavily prescribed in writing (nominative case, avoidance of contraction, and *who*).

two-level factors. Model comparison was conducted via asymptotic likelihood ratio tests between a model without interactions, one adding the two-way interactions, and one additionally including the three-way interaction. Maximal random effect structure was used when supported by the data (as recommended by Barr et al. 2013): the random effects were successively simplified, chosen by lowest variance, until convergence was reached. The final model included a random intercept for pronoun gender. Adding the two-way interactions

Table 3: Twitter and iWeb comparison: model fixed effects. Each factor is reported by one of its two complementary variants (e.g. "Nominative" as one of the instances of case), and the estimates describe an increase (or decrease) on a logit scale.

Effect	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	1.09	0.29	3.75	<0.001	***
Nominative	−0.74	0.30	−2.44	0.015	*
Uncontracted	−1.75	0.22	−7.79	<0.001	***
Who	0.13	0.19	0.72	0.47	
Nominative:Uncontracted	−0.44	0.25	−1.73	0.08	
Uncontracted:Who	−0.26	0.28	−0.93	0.35	
Nominative:Who	−0.72	0.27	−2.61	0.009	**

Significance thresholds: $^{***}p < 0.001$, $^{**}p < 0.001$, $^{*}p < 0.001$.

improved the fit ($p < 0.001$), while the three-way interaction did not ($p = 0.45$). The model results are reported in Table 3.

The model confirms that in Twitter clefts use fewer nominative pronouns ($p = 0.015$) and are more frequently contracted ($p < 0.001$). An interaction shows that clefts in which nominative pronouns and *who* co-occur appear less in Twitter than in iWeb ($p = 0.009$).

4 Results

The results from the analysis in Section 3.2 confirm some of the hypotheses outlined in the introduction and in Section 2.3. The pattern that Cheshire et al. (2013) noted on relative clauses, whereby the complementiser *who* appeared most frequently with subjects, was replicated in it-clefts as the only association that the syntactic role of the clefted element showed: when an object was clefted, the complementiser *that* was used in almost three quarters of cases, while a clefted subject appeared significantly more with *who* as a complementiser (see Figure 6).

While it is still not clear whether the case of the pronoun, and more specifically the nominative case, has a direct bearing on emphasis (as Maier 2014 suggests), an association was found between nominative pronouns and the presence of emphasis markers within the cleft (see Figure 3). This could mean either that two emphatic features are used concomitantly, or that the association is due to other features of the sentence that we are not able to see. Nominative case was moreover associated with uncontracted clefts, which rank high in the formality axis (while contraction has no effect on fuzziness and therefore does not directly influence emphasis).

Our prediction in Section 2.3 that the complementiser *who* may be associated with the avoidance of contraction, stemming from the fact that this form ranked higher in formality and expressiveness, was confirmed (see Figure 4). This complementiser was likewise associated with nominative pronouns, which, however, cannot be explained as a less fuzzy version of accusative case pronouns. The hypothesis that the presence of emphasis markers could be associated with uncontracted clefts or *who* was not borne out.

The difference outlined between Twitter and general-purpose internet English shows how social media language is more variable than edited writing. This is in keeping with the "conceptual orality" interpretation of social media language. The sources in the iWeb corpus show very stark tendencies, which may be due to the editorial process its texts undergo: most notably, the avoidance of contraction is almost ubiquitous (93.7%), and it is known to be one of the most prescribed-for traits of written language.

5 Discussion

Our main data on the variability of English it-clefts hails from written texts on the social media platform Twitter, which we compare to general purpose written English web texts. We start from the situation that one of the variants is heavily preferred prescriptively in English writing, as confirmed by the edited web data. Contractions are typically avoided in written English (e.g. Wikipedia contributors 2019: List of English contractions).

Most of the current grammars of English discussing cleft sentences do not focus on the case of the pronoun, or on the distribution of the different form variants of the features we selected (e.g. Aarts 2011; Hilpert 2014; Cummings 2018; Depraetere and Langford 2019). Even though some mention that accusative forms are common in non-structural positions such as clefts (Cambridge Dictionary 2019; Eastwood 1994: p. 62), most grammar advice in schools and online still prescribes nominative case for this construction (cf. Grammar Book 2019; Quirk et al. 1985: p. 338 calls the accusative form "informal"). An exception to this are the Longman Grammar (Biber et al. 1999: p. 336), where both nominative and accusative forms are seen as possible but nominative pronouns are noted to be "presumably felt to be more correct" in formal registers, and the Cambridge Grammar (Huddleston and Pullum 2008: p. 1414ff), which characterises the nominative as "formal

or very formal" and the accusative as preferred in colloquial uses. Similarly, Maier (2013) shows that while nominatives dominate in written English corpora for clefts, spoken corpora exhibit many more accusatives.

As for the complementiser, most grammars mention both *who* or *that* for human referents and/or subjects, with *that* used in other cases (Eastwood 1994: p. 62). The Longman Grammar (Biber et al. 1999) notes that in written registers, *who* is often chosen, while *that* or the omission of the complementiser are more typical of casual speech.

The prescriptively "correct", or at least preferable, version of the English it-cleft thus seems to be uncontracted, using a nominative pronoun, and (given that most clefted elements are human subjects), using the complementiser *who* (though *that* is also possible, in particular for objects).

Looking at formality and emphasis more in detail (as discussed in Section 2.2–Section 2.3), more formal texts generally avoid contractions, and, as noted above, tend to favour the nominative pronoun in the context of cleft. As for the complementiser, it is not immediately clear which one is more formal. However, following the definition of formality in Heylighen and Dewaele (1999), more precise word forms are more formal than general-purpose words, because they avoid semantic fuzziness and context-dependence. Under this view, *who* is more formal than *that* when applied to human referents (as in all of our examples), since it is more precise in selecting specifically for humans.

The formal variants thus overlap with the prescriptively preferred variants. And indeed, the iWeb corpus of edited online written English shows the formal/prescriptive version of the it-cleft almost categorically, as shown in Figure 7: each formal feature is chosen in about 90% of the instances. This exemplifies a tendency that writers have, even in online texts, to adhere to formal style standards. The Twitter data is even more interesting in this light, since it shows considerable variation between the formal and informal variants.

As Heylighen and Dewaele (1999) note, expressivity/emphasis is somewhat orthogonal to (in)formality, and higher emphasis is associated with higher specificity but also higher context-dependence. Note that uncontracted clefts are able to receive more stress than contracted ones, and thus can be more emphatic. As argued above, *who* is more specific than *that*, thus also more emphatic under the assumed definition. Finally, Maier (2014) claims that nominative pronouns are more emphatic than accusative ones. This does not follow in Heylighen and Dewaele (1999)'s framework, as nominative pronouns are not more semantically specific or context-dependent than accusative ones. Here, we show that the case of the clefted pronoun in Twitter is not used to mark the syntactic role of the referent, which may free up the case feature to mark something else.

Though written, Twitter data tends towards a very informal style by default, exhibiting many features typical of conversational spoken language, and many informal variants (Koch and Oesterreicher 1985; Storrer 2013; Eisenstein 2013b; Hu et al. 2013). In Twitter, using formal style features constitutes a marked case that, by being somewhat unusual, increases the general emphasis on the utterance. Taken together, we believe that the dimensions of (in)formality and emphasis at least in part explain the "clustered" distribution of cleft variants seen in Figure 2. First, we observe that the lack of enforcement of prescriptive standards leads to a much higher variability in this medium than other written (web) text. Second, the formal variant that dominates in edited text (Uncontr-NOM-who) is still highly present in the Twitter data, but it is joined equally by two other variants. Of these, one reflects informal usage, which is default on Twitter (Contr-ACC-that). Further, the more specific complementiser *who* is also used often in informal Twitter clefts, increasing the emphasis in these cases (without seeming too formal, since both *who* and *that* are considered standard for clefts). Finally, further emphasis may be added by using the nominative pronoun and/or by adding orthographic emphasis markers, which are associated with nominative pronouns. Other combinations are rarer, possibly because of a resulting clash between the underlying utterance properties of (in)formality and emphasis.

In summary, in this paper we studied English it-clefts in written social media texts. We observe considerable variation in the realisation of the clefts with regard to the (non)contraction of the copula, the case of the pronoun, and the complementiser used. This variation is not random – instead we observe several close associations of the studied features, as well as the presence of additional orthographic emphasis markers. We conclude that while general web text is subject to overt standard language conventions, these are less relevant in Twitter and uncover different subtypes of it-clefts in form. We characterise these subtypes in terms of their placement on scales of formality and emphasis. It would be very interesting to see if the variation in form also

corresponds to further functional distinctions among these it-clefts (for example, as far as their information structure or the relative salience of clefted referents in subsequent discourse). We must leave these questions for future research.

Acknowledgment: The authors would like to thank the anonymous reviewers for their detailed comments. We are grateful to our project collaborators Berfin Aktaş, Yulia Clausen, and Manfred Stede, as well as Heike Pichler and Joseph DeVeaugh-Geiss for discussions about this work. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 317633480 – SFB 1287, Project A03/MGK.

Appendix I Twitter search queries

This appendix lists the queries used to mine Twitter for cleft sentences through TAGS (Hawksey 2016):

- Lines 1–8 search for accusative pronouns, 9–16 for nominative pronouns;
- Lines 1–4 and 9–12 search for *that*, 5–8 and 13–16 for *who*;
- Lines 1, 2, 5, 6, 9, 10, 13 and 14 search for uncontracted clefts, the others for contracted clefts;
- Odd lines search for feminine pronouns, even lines for masculine pronouns.

1. `%22it%20is%20her%20that%22 AND -filter:retweets`
2. `%22it%20is%20him%20that%22 AND -filter:retweets`
3. `%22it%27s%20her%20that%22 AND -filter:retweets`
4. `%22it%27s%20him%20that%22 AND -filter:retweets`
5. `%22it%20is%20her%20who%22 AND -filter:retweets`
6. `%22it%20is%20him%20who%22 AND -filter:retweets`
7. `%22it%27s%20her%20who%22 AND -filter:retweets`
8. `%22it%27s%20him%20who%22 AND -filter:retweets`
9. `%22it%20is%20she%20that%22 AND -filter:retweets`
10. `%22it%20is%20he%20that%22 AND -filter:retweets`
11. `%22it%27s%20she%20that%22 AND -filter:retweets`
12. `%22it%27s%20he%20that%22 AND -filter:retweets`
13. `%22it%20is%20she%20who%22 AND -filter:retweets`
14. `%22it%20is%20he%20who%22 AND -filter:retweets`
15. `%22it%27s%20she%20who%22 AND -filter:retweets`
16. `%22it%27s%20he%20who%22 AND -filter:retweets`

Appendix II Raw data numbers

Table 4: Raw numbers of the feature variants in the two Twitter data collections.

First round of data collection	
Case	
Nominative	311
Accusative	487
Contraction	
Contracted	485
Uncontracted	313

Table 4: (continued)

First round of data collection	
Complementiser	
That	286
Who	512
Referents no.	
One	650
Two	148
Clefted Subj	112
Clefted Obj	36
Emphasis markers	
No Emphasis	685
Emphasis	113
Exclamation	55
Capital pronoun	33
Capital pro. + !	3
Capital initial	4
Capital cleft	2
Asterisks	4
Nonstandard comma	6
Nonst. comma + !	1
Other	5
Total number of clefts	798
Second round of data collection	
Case	
Nominative	242
Accusative	374
Contraction	
Contracted	405
Uncontracted	211
Complementiser	
That	256
Who	360
Total number of clefts	616

References

Aarts, Bas. 2011. *Oxford modern English grammar*. Oxford: Oxford University Press.

Akmajian, Adrian. 1970. On deriving cleft sentences from pseudo-cleft sentences. *Linguistic Inquiry* 1(2). 149–168.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 1–44.

Bates, Douglas, Martin Mächler, Bolker Ben & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Bouma, Gerlof, Lilja Øvrelid & Jonas Kuhn. 2010. Towards a large parallel corpus of clefts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Valetta: Malta.

Cambridge Dictionary. 2019. *Grammar*. Available at: https://dictionary.cambridge.org/grammar/british-grammar/pronouns-personal-i-me-you-him-it-they-etc (accessed 4 May 2020).

Cheshire, Jenny, David Adger & Sue Fox. 2013. Relative *who* and the actuation problem. *Lingua* 126. 51–77.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edn. Abingdon-on-Thames: Routledge.

Collins, Peter C. 1991. *Cleft and pseudo-cleft constructions in English*. London & New York: Routledge.

Cummings, Louise. 2018. *Working with English grammar: An introduction*. Cambridge: Cambridge University Press.

Davies, Mark. 2018. *The 14 billion word iWeb corpus*. Available at: https://www.english-corpora.org/iweb/ (accessed 4 May 2020).

Delin, Judy. 1992. Properties of it-cleft presupposition. *Journal of Semantics* 9(4). 289–306.
Depraetere, Ilse & Chad Langford. 2019. *Advanced English grammar: A linguistic approach*. London: Bloomsbury.
Eastwood, John. 1994. *Oxford guide to English grammar*. Oxford & New York: Oxford University Press.
Eisenstein, Jacob. 2013a. Phonological factors in social media writing. In *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, 11–19. Atlanta, Georgia.
Eisenstein, Jacob. 2013b. What to do about bad language on the internet. In *Proceedings of HLT-NAACL*, 359–369. Available at: http://www.aclweb.org/anthology/N13-1037.
Grammar Book. 2019. Available at: https://www.grammarbook.com/grammar/pronoun.asp (accessed 4 May 2020).
Hawksey, Martin. 2016. TAGS: Twitter archiving Google spreadsheet. v. 6.1. Available at: https://tags.hawksey.info/ (accessed 4 May 2020).
Heath, Maria. 2018. Orthography in social media: Pragmatic and prosodic interpretations of caps lock. *Proceedings of the Linguistic Society of America* 3(55). 1–13.
Heylighen, Francis & Jean-Marc Dewaele. 1999. *Formality of language: Definition, measurement and behavioral determinants*. Internal Report, Center "Leo Apostel", Free University of Brussels.
Hilpert, Martin. 2014. *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.
Hope, Adery C. A. 1968. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society* 30. 582–598.
Hu, Yuheng, Kartik Talamadupula, Subbarao Kambhampati. 2013. Dude, srsly? The surprisingly formal nature of Twitter's language. In *Proceedings of the International Conference on Weblogs and Social Media*. Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6139 (Accessed 4 May 2020).
Huddleston, Rodney & Geoffrey K. Pullum. 2008. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Kjellmer, Goran. 1997. On contraction in modern English. *Studia Neophilologica* 69. 155–186.
Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe—Sprache der Distanz. *Romanistisches Jahrbuch* 36. 15–43.
Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
Lambrecht, Knud. 2000. When subjects behave like objects: An analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language* 24(3). 611–682.
Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3). 3–463.
Maier, Georg. 2013. As the case may be: A corpus-based approach to pronoun case variation in subject predicative complements in British and American English. In Magnus Huber & Joybrato Mukherjee (eds.), *Corpus linguistics and variation in English: Focus on non-native Englishes*, vol. 13. *Studies in variation, contacts and change in English*. Helsinki: VARIENG. Available at: http://www.helsinki.fi/varieng/series/volumes/13/maier/ (accessed 4 May 2020).
Maier, Georg. 2014. The case of focus. In K. Davidse, et al. (ed.), *Corpus interrogation and grammatical patterns*, 173–205. Philadelphia: John Benjamins.
McAteer, Erica. 1992. Typeface emphasis and information focus in written language. *Applied Cognitive Psychology* 6. 345–359.
Nelson, Gerald, Sean Wallis & Bas Aarts. 1998. *International Corpus of English – Great Britain (ICE-GB) release 2*. Available at: https://www.ucl.ac.uk/english-usage/projects/ice-gb/.
Patten, Amanda L. 2010. *Cleft sentences, construction grammar and grammaticalization*. University of Edinburgh Doctoral dissertation. Available at: https://era.ed.ac.uk/handle/1842/27175.
Patten, Amanda L. 2012. *The English it-cleft. A constructional account and a diachronic investigation*. Berlin and Boston: De Gruyter Mouton.
Prince, Ellen F. 1978. A comparison of wh-clefts and it-clefts in discourse. *Language* 54(4). 883–906.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Svartvik Jan. 1985. *A comprehensive grammar of the English language*. London & New York: Longman.
Scheffler, Tatjana. 2017. Conversations on Twitter. In Darja Fišer & Michael Beißwenger (eds.), *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world Book series Translation Studies and Applied Linguistics*, 124–144. Ljubljana: Ljubljana University Press.
Scott, Kate. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics* 81. 8–20.
Storrer, Angelika. 2013. Sprachstil und Sprachvariation in sozialen Netzwerken. In Barbara Frank-Job, Alexander Mehler & Tilmann Sutter (eds.), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften.
Tagliamonte, Sali A. & Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1). 3–34.
Wikipedia contributors. 2019. *Manual of style — Wikipedia, the free encyclopedia*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/ (accessed 4 May 2020).
Zimmermann, Malte. 2008. Contrastive focus and emphasis. *Acta Linguistica Hungarica* 55(3–4). 347–360.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2019-0066).