We Who Tweet: Pronominal Relative Clauses on Twitter

Kirby Conrod, Rachael Tatman and **Rik Koncel-Kedziorski**
Department of Linguistics
University of Washington
kconrod@uw.edu, rctatman@uw.edu, kedzior@uw.edu

Abstract

Pronominal relative clauses were previously reported to be unproductive in English, appearing only in Bible verses and proverbs. This corpus study of Twitter data shows that pronominal relative clauses are a productive part of contemporary English, and can be used in both literary and nonliterary registers.

1 Introduction

Pronominal relative clauses, also known as Voldemort phrases (Zobel, 2015), are relative clauses headed by pronouns. Pronominal relative clauses (PRCs) can be restrictive (PRRCs) or non-restrictive. In previous works that have studied PRCs it has been claimed that the construction is not productive, but is only found in particularly literary contexts (Curme, 1912; Elbourne, 2013; Zobel, 2015). In order to better determine whether PRCs are a productive part of modern English syntax, and to test empirical generalizations that have been made about the construction, we collected and analyzed data from Twitter which contains PRCs.

PRRCs are pronominal relative clauses which have a restrictive reading available. In the Twitter data that we have collected, we found many examples of PRRCs:

(1) He who has great power should use it lightly. [twi.82]

(2) We who #FeelTheBern feel the same about you! [twi.7511]

The data we collected included many PRRCs of a 'literary' style–either Bible verses, quotes from famous politicians, or old adages that may be at this point idiomatic. (1) above is one such 'literary' PRRC, in this case a translation of a quote from Seneca, a Roman Stoic philosopher. In contrast, (2) is a completely non-literary example of a PRRC, and cannot be traced to any religious or idiomatic history.

Non-restrictive pronominal relative clauses, by contrast, are appositive relative clauses headed by referential pronouns:

(3) I can do all things in him who strengthens me. [twi.4746]

(4) How is it that he, who showed a CLOCK to an ENGINEERING teacher got arrested? [twi.727]

We also found both literary and non-literary uses of this non-restrictive version throughout the Twitter data. We used repetition and uniqueness to encode literariness. Our data shows that PRCs are a productive part of contemporary English. Through further analysis of PRRCs we estimate that a PRRC is tweeted every thirty seconds. Additionally, of the PRCs that we sampled, 37% (1222 tweets) were unique and non-literary.

2 Background of Pronominal Relative Clauses

Elbourne's (2013) syntactic analysis of PRRCs (Voldemort phrases) proposes a structure in which the pronominal head is a definite determiner. The analysis of relative clauses in Elbourne (2013) is agnostic as to whether the pronominal head originates within the relative clause or merges externally. Elbourne (2013) does include a nominal layer in the structure that constitutes a null element meaning something like 'person.'

(5) [[he[*person* [who. . .]]]]

(Elbourne, 2013)

Elbourne (2013) does not propose a structural difference between restrictive and non-restrictive relative clauses.

In a recent semantic analysis, Zobel (2015) proposes that PRRCs constitute generic statements about generic kinds: that is, that PRCs such as *he who prays* are equivalent to *the kind of man who prays*. The insight of this semantic derivation captures what makes pronouns–which otherwise would be referential expressions–allowable as heads of restrictive relative clauses. In order to head a restrictive relative clause, a pronoun must not be referential, must not have an antecedent, and must not be interpreted as specific. Zobel (2015) derives a semantic structure of PRRCs that reduces down to generic kinds: *he* is interpreted as something more like *the sort of man*. This is well in keeping with the syntactic definiteness of the pronoun.

In these two analyses of PRRCs and an oft-cited earlier description by Curme (1912), however, several generalizations have been made about PRRCs without robust support. Elbourne (2013), Zobel (2015), and Curme (1912) all work with the assumption (citing Curme's (1912) generalization) that PRRCs are not a fully productive construction in contemporary English. Following this assumption, examples that (Zobel, 2015) and (Elbourne, 2013) use are Bible quotes, proverbs, or literary quotes:

(6) He who abides in love abides in God.

((Zobel, 2015): 7)

(1 John 4:16 NKJV)

(7) He who hesitates is lost.

((Elbourne, 2013): 205)

In keeping with these assumptions, examples that (Elbourne, 2013) and (Zobel, 2015) use are themselves proverbs and quotes. This study provides evidence to that the construction is not in fact limited to Biblical language or literary reference.

3 Methods

Twitter data was extracted from the Twitter public application programming interface (API) with an R (R Core Team, 2015) script using the plyr (Wickham, 2011) and twitteR (Gentry, 2015) libraries[1]. Up to the 1000 most recent tweets (on September 22, 2015) were collected for ten exact phrase *PRO who* for each pronoun.

To facilitate hand-analysis, we applied a cascade of strategic filters to remove unwanted tweets (i.e. non-PRCs) from our data. For each we provide a description of the filter and its intended targets below. Many of these filters are designed to discard specific syntactic phenomena that would be more readily recognizable given a parse structure. However, due to the casual nature of twitter orthography, accurate automatic syntactic parsing is rarely available. We use simple, strategic filters to approximately characterize undesired constructions. They are designed with a bias toward removing too few rather than too many tweets.

First, we apply exact string match to remove duplicates. From the remaining tweets we remove those where the last non-whitespace character before the word *who* is anything but a letter or a comma. This also excludes utterance-initial *who*. We then remove tweets with the pattern *of PRO who* (where *PRO* varies over English pronouns) as these are typically partitives rather than PRRCs. We then filter tweets where *it* precedes *who* by 2 to 4 words, e.g *It is we who* or *It may have been we who*. These cleft constructions are not PRCs. Finally, we remove tweets where the *PRO who* combination is preceded by a clause taking verb from among the inflected forms of *ask, tell, wonder, inform,* and *show*. This filtered out tweets with clausal complements, e.g. *show me who did it*.

This filtering process left us with 3261 tweets of the original 10,000. We sampled 300 of these for detailed hand analysis.

We cast the remaining 3261 tweets into two classes, based upon the uniqueness of the words immediately following *who*. If the two word sequence following *who* is unique, we consider this evidence that it is a non-literary use. This non-literary class consists of 1222 examples.

The data was hand-tagged for availability of a restrictive reading, availability of an appositive reading, the head of the relative clause, the case of the pronominal head, and what role the relative clause played in the matrix clause.

Restrictive and appositive readings were not taken to be mutually exclusive. A relative clause was tagged as restrictive if there was an interpretation available where the relative clause denoted some subset of a (usually generic) set of entities denoted by the head. A relative clause was tagged

[1] All code and data used will be made available.

as appositive if there was an interpretation available where the relative clause denoted a property held by the entire set of entities denoted by the head. Relative clauses could be tagged as both restrictive and appositive if both readings were available; no relative clauses were tagged as neither restrictive nor appositive.

Heads were tagged simply as the pronouns *he, him, she, her, they, them, I, me, we, us,* and *you*. The case of the head was tagged as nominative (NOM) if it was tagged as *he, she, they, I, we*, and accusative (ACC) if it was *him, her, them, me,* or *us*. We did not tag *you* as either NOM or ACC because there is no overt morphological difference between the two forms.

The role of the relative clause in the matrix clause was tagged S if the RC was a subject, O if the RC was an object, P if the RC was a predicate, and F if the RC was a fragment, or did not play a role in a matrix clause.

4 Results

4.1 Literariness and Uniqueness

The data we collected allowed us to address several generalizations made about PRCs. One restriction Zobel (2015) places on PRCs is based on Curme's same claim (1912): that PRCs are archaic, and not productive in English– and that they appear only in Bible texts, proverbs, and other sayings. This is a claim not about the structure of PRCs, but about their historicity, register, and productivity as a syntactic construction. Firstly, Zobel's (2015) and Elbourne's (2013) ability to produce novel examples of PRCs for syntactic and semantic analysis indicates that speakers of English retain the ability to create novel sentences using this syntactic structure. Secondly, 37% of the tweets containing PRRCs that we sampled (1222 tweets) were unique, which we take as evidence that they are non-literary.

Our method for determining literariness, while less ideal than comparing against a large corpus of literature, is quite conservative and has the advantage of capturing facts about contemporary usage. For instance, noting that many PRCs deal with biblical topics, we searched the King James Version of the Bible (KJV) for the 4-grams *PRO who* w_1w_2, where *PRO* is some pronoun and w_1w_2 are words that follow *who* in one or more of the 3261 filtered tweets. This resulted in a mere 227 matches. However, there are more than 227 Biblical tweets which are considered literary by our metric. This is due in part to the fact that phrases from the bible whose original context is non-PRRC are often used in PRRC constructions (for example, *PRO who comes in the name of the Lord* does not appear in the KJV, but does appear frequently in our twitter data).

By using 4-grams such as *PRO who* w_1w_2 to test for repetition, we labelled as literary exact repetitions of phrases as well as near-repetitions that use templatic patterns for the purpose of literary allusion. This is an inclusive measure of literariness that labels more tweets as literary than those that exactly correspond to texts such as the KJV Bible; our hope is that the tweets remaining tagged as non-literary should be more reliably unique. Both measures of uniqueness versus literariness–the comparisons within our own data set and the comparisons to the KJV Bible–indicate that pronominal relative clauses are not always either literary or Biblical.

4.2 Possible Pronominal Heads

While this is not central to her semantic analysis, Zobel (2015) also claims that PRRCs are only headed by masculine pronouns, while other (non-*he*) pronominal RCs are non-restrictive. Contrary to this claim, we found many instances of PRRCs headed by the full range of English pronouns, including *he*, *she*, *him*, *her*, *they*, *them*, *we*, *us*, and *you*.

(8) But value he who shows you respect, honesty & trust. [twi.310]

(9) A preacher that fears the powers that are contemporary, and dismisses the power of him who is eternally in power, is not fit to lead people [twi.4673]

(10) She who leads rules, so play nice or I won't let you [twi.1455]

(11) Every moment is a golden one for him/her who has the vision to recognize it as such. [twi.5951]

(12) they who control the pumpkin spice control the universe [twi.2610]

(13) It's a funny old game ain't it but them who take part in it wouldn't change it [twi.6439]

(14) And we who know and realize this should always be willing and eager to save others and not condemn them [twi.7637]

Figure 1: Proportion of PRRC's with each head by literariness.

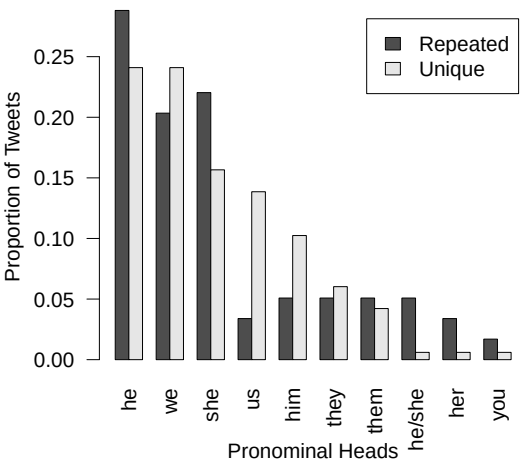
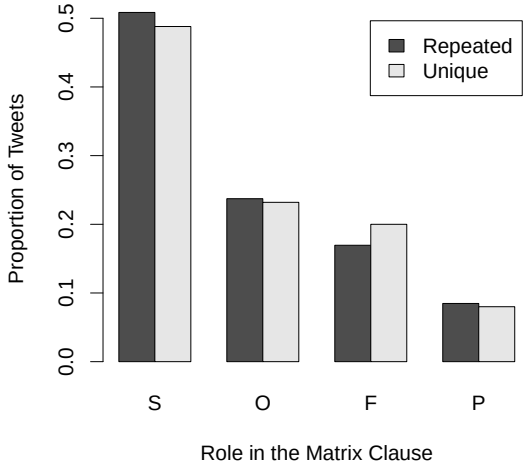


Figure 2: Proportion of PRRC's in each role in the matrix sentence.



(15) u can't tease us who weren't there with a new song and not let us hear!!!! [twi.9453]

(16) you who make me smile, you are what makes my heart [twi.8765]

Of the restrictive constructions in the data we analysed, the only pronominal head we did not find was *it.* While there is a clear preference for *he,* this is by no means a universal constraint, and so cannot be relied upon to provide the semantic denotation derived by PRRCs. In addition, this preference is somewhat weaker when looking only at the unique tweets, as can be seen in Figure 1. (Although it was not statistically significant at α = 0.05: χ^2(9, N= 225) = 15.44, p = 0.079.)

4.3 Role in the Matrix Clause

Zobel (2015) states that PRRCs "combine with object-level predicates" at the matrix level: that is, the PRRC is usually at the left periphery, or is the subject of, the matrix sentence. To test whether this was the case, we separated out unique and repeated PRRCs and analyzed their syntactic role in the matrix sentence. Each tweet was tagged with a matrix syntactic role: S (subject), O (object), P (predicate), or F (fragment). Among all PRRC's the subject position was the most common, account for 49% of PRRCs in the data-set, object position (23%) and fragments (19%) were also quite common. Predicate PRRC's (8%) were by far the rarest. These proportions held across both types–unique and repeated. This is summarized in Figure 2. This suggests that there's more diversity in these construction than has previously been posited.

5 Restrictives and Appositives

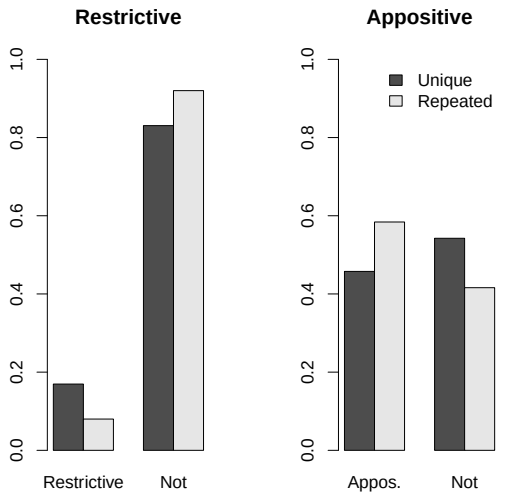
Restrictive and appositive tweets were also analyzed by uniqueness. There was no statistically reliable difference between unique and repeated tweets in their use of either restrictive or appositive structures (χ^2(1, N= 184) = 2.45, p = 0.117 and χ^2(1, N= 184) = 2.09, p = 0.147, respectively). This is summarized in Figure 3. As can be seen in the figure, roughly the same proportion of unique and repeated tweets were tagged as restrictive or appositive.

6 Conclusions

Based on the data collected, we conclude that pronominal restrictive relative clauses are a productive and robust part of contemporary English. Based on our data, we estimate that a PRRC is tweeted every thirty seconds. Additionally, of the PRCs that we sampled, 37% (1222 tweets) were unique and thus, we argue, non-literary. This is evidence that, while they may be stylistically marked in some way, PRCs are a construction that is productively available to contemporary English speakers.

Further, there are not predictable differences between repeated and unique PRC's, which seems to suggest that productive and non-productive uses

Figure 3: Proportion of unique and repeated clauses that are restrictive and appositive.



are making use of the same underlying structures. The most striking difference between repeated and unique tweets was that the latter were more likely to be headed by pronouns other than *he* or *she*– which is a semantic rather than syntactic distinction.

References

George O Curme. 1912. A history of english relative constructions. *The Journal of English and Germanic Philology*, 11(3):355–380.

Paul Elbourne. 2013. *Definite descriptions*. OUP Oxford.

Jeff Gentry, 2015. *twitteR: R Based Twitter Client*. R package version 1.1.9.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Hadley Wickham. 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Sarah Zobel. 2015. Voldemort phrases in generic sentences. *Grazer Linguistische Studien*.