Measuring Generic Structures in an English Novel

Seminar Paper
Submitted to Prof. Ariel Cohen
By Gal Abramovitz
September 2019

1 Abstract

As generic sentences appear to derive from non-generic (i.e. episodic) sentences (Carlson and Pelletier 1995), the grammatical structures that are used to express genericity are not exclusive to generics (Dahl 1995).

These structures that allow manifestation of generic NPs are generally restricted to bare plural nouns, definite singular nouns or indefinite singular nouns, with bare plurals being the most common structure in modern English (Cohen 2002).

While the above claim seems well-accepted, the complementing question - regarding the commonality of these grammatical structures' generic interpretation - remained unattended.

Using George Orwell's novel Nineteen Eighty-Four as a corpus, I have tried to address both the probability of a bare plural being interpreted generically and the semantic and grammatical conditions that encourage these generic readings.

2 Methodology

I have decided to take an empirical approach in my research and to base my conclusions strictly on measured data. Hence, I have manually sorted all plural nouns occurrences in the corpus, separating the BPs from non-BPs. Then I tagged the BP occurrences that that have a salient generic interpretation.

This tagged bare plurals database allowed me to primarily answer my research question, but more importantly to analyze the sentences in which BPs were interpreted generically and find common characteristics among them.

3 Definitions

Before tagging BPs in general and generic BPs in particular, I had to explicitly define what I was looking for.

While the term “bare plurals” is pretty self-explanatory - plural nouns that aren’t preceded by determiners - the definition for generic sentences is much more elusive: many sentences had the generic “feel”, although they weren’t meeting the formal guidelines for generics. For example, consider the following sentences:

(1) a. Physical facts could not be ignored

b. Physical facts are not ignorable

Even though (1.a) and (1.b) “feel” equivalent, the sentence that appears in the corpus (1.a) contains an overt modal quantifier, which disqualifies it as a generic sentence.

Nevertheless, in order to have some guidelines to filter out non-generic BPs, I used various rules of thumb, as exemplified in the following non-generic sentences from the corpus (with the relevant BPs underlined):

(2) “Oranges and lemons”, say the bells of St. Clement's.

(3) Everyone kept asking you for razor blades.

(4) People with dark hair sometimes had blue eyes.

- The BP must occur in a *proper* sentence (i.e. with a VP).

  For example, the quote *"oranges and lemons"* in (2) isn't a proper sentence, hence these aren't generic BPs.

- The relevant BP must be in the subject position in the discussed sentence.

  For example, in (3) the BP *razor blades* is the sentence's object rather than its subject, thus the sentence isn't even a candidate as a generic sentence with this BP.

- The VP describing the BP must not be described with a frequency adverb.

  For example, in (4) the adverb *sometimes* makes the sentence non-generic.

These definitions were used in order to roughly *rule out* some BPs occurrences as generics, rather than to *identify* the generic sentences.

4 Findings

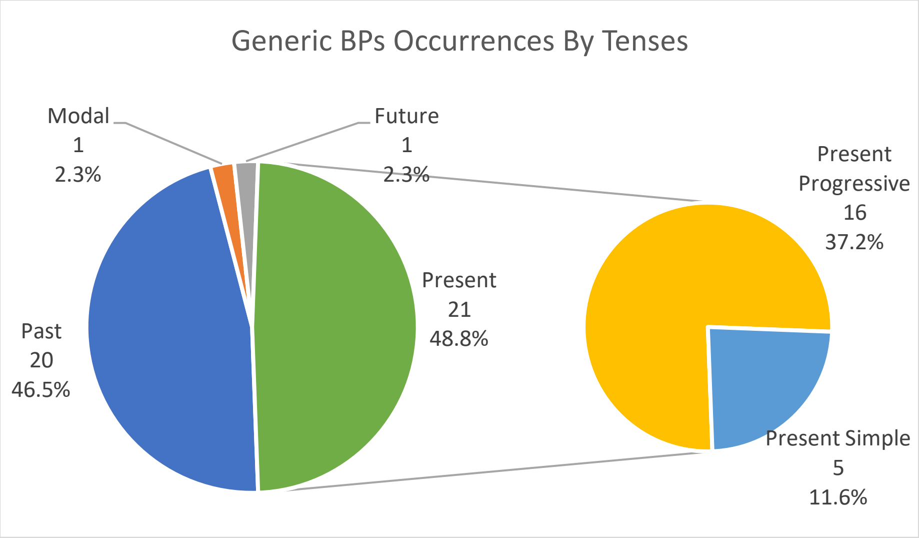
The novel Nineteen Eighty-Four contains 4102 plural nouns. Manual examination of these plural nouns resulted in a count of 40% bare plurals (1627 listings).

Essentially, I have found only 2.64% of the sentences containing bare plurals to have a salient generic interpretation (43 listings).

The immediate conclusion is that even though bare plurals seem to be the most common manifestation of generics in English, most bare plurals are not generics. A possible implication of this conclusion is that it's plausible and relatively simpler to find contexts in which generic BPs are *unlikely* to occur, rather than characterizing the contexts in which they do appear.

Further examination of the tagged sentences resulted in the following generalizations regarding the contexts in which generic BPs occur:

1. 95.3% of all generic BPs have occurred in the present and the past tenses and were divided almost equally between them.



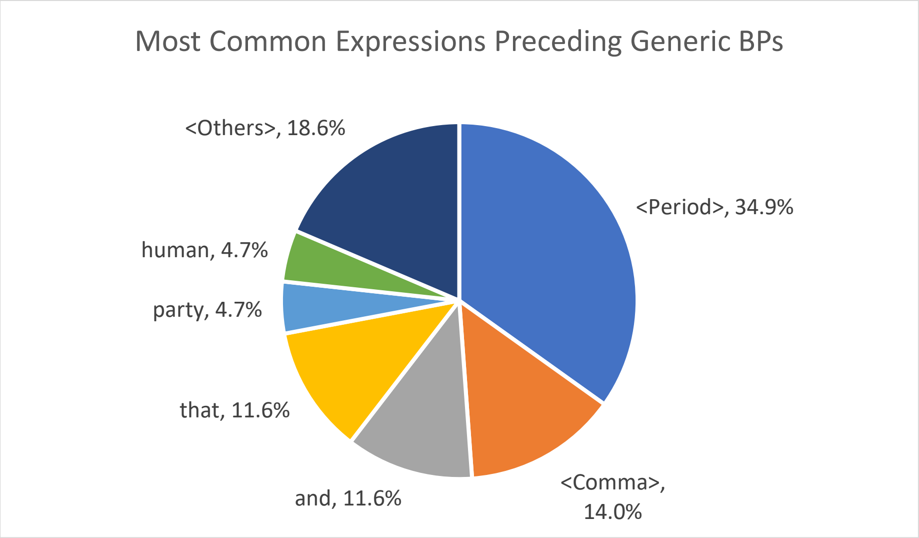
2. As mentioned by Carlson (2010), in a preliminary study by Frazier and Clifton they reasoned that t-properties are more likely to be interpreted generically in the past tense, while k-properties would usually have a more salient non-generic interpretation.

   This paper's findings support their hypothesis, as in fact 85% of the tagged generic sentences in past tense described t-properties. Having said that, this research didn't include tagging non-generic k-properties sentences in the past tense, hence the claim isn't fully backed up by the data that was collected.

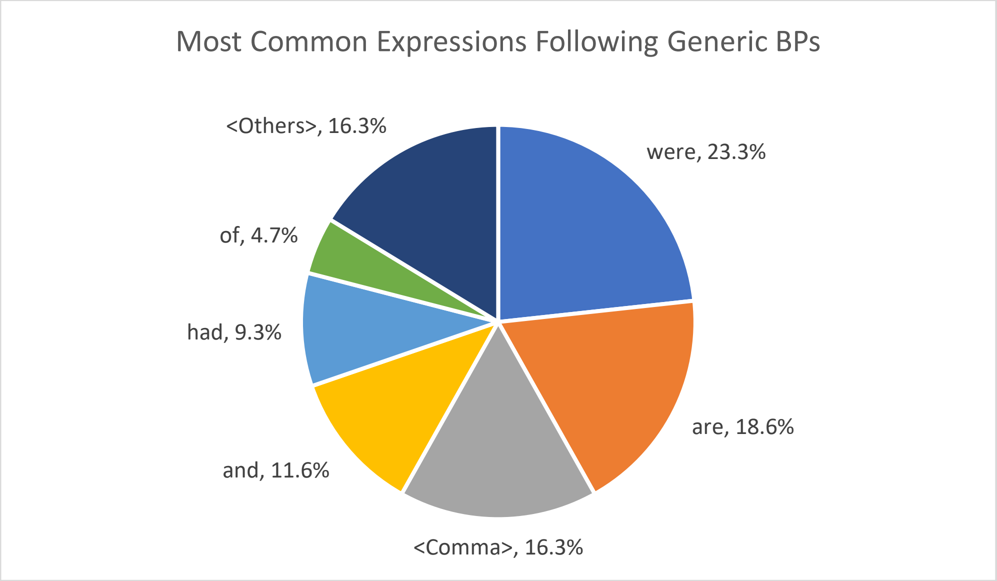
3. By examining the neighboring words of the generic BPs, I was hoping to find structures that are more likely to contain them:

a. 39.5% of generic BPs occurred right in the beginning of sentences: following a period, a semicolon or a colon that was followed by a quotation mark.

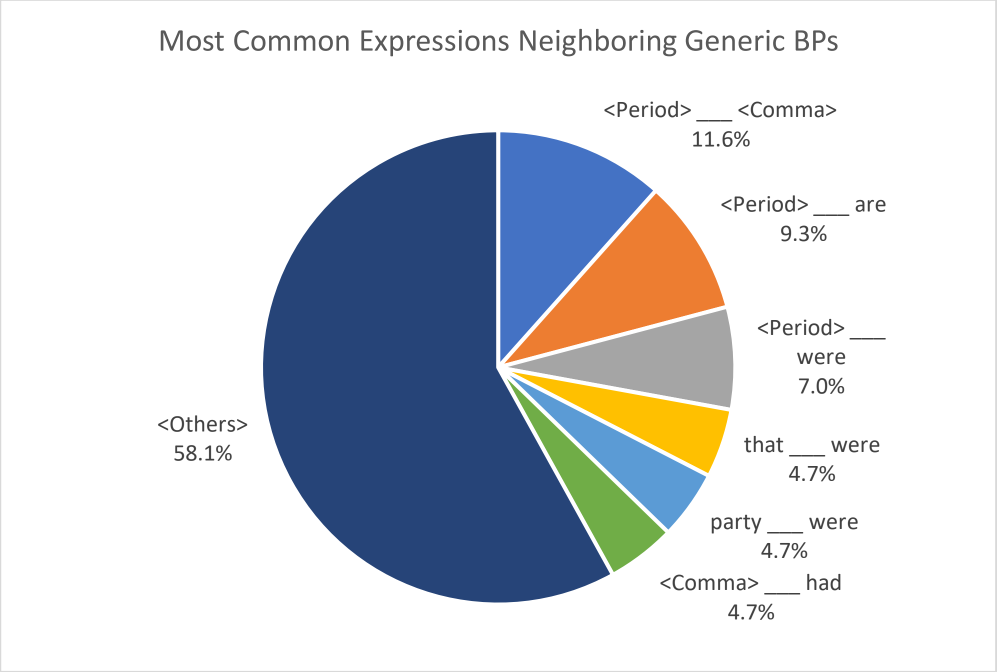
These results indicate that generic BPs are likely to appear without adjectives describing them, as English adjectives precede the noun they're describing.



b. Predictably, 41.9% of generic BPs were followed by the auxiliary verbs *are* or *were*.



c. There were no distinctly common structures that included both neighboring expressions.



Note: The <Others> labels denote expressions that appeared only once in the data.

4. BPs from the semantic fields of time, quantification and measurements (e.g. *hundreds, years, fifties, kilometres*) tend to have non-generic interpretations.

   Out of 37 words that were picked by this category and have occurred in a total of 542 sentences, **none** occurred in a generic context.

5 Validity

Considering that the tagging of both BPs and generic BPs was done manually, I humbly hope my judgement and knowledge of generics have contributed my efforts to keep the recall and precision rates as high as possible.

In order to assemble the initial list of sentences containing plural nouns I've used the NLTK (i.e. Natural Language Tool Kit) Python package. It has tagged plural nouns in 88.5% of the sentences in the corpus, which means the maximal recall rate is bound by this percentage, as I have read and checked each tagged sentence manually.

Regarding the precision rate - during the manual tagging, I had noticed that some of the allegedly plural nouns weren't actually plural. This mainly occurred with names and a few Present Simple verbs with the *s* suffix. Additionally, the only plural noun that NLTK didn't tag, and I was aware of, was the Latin-originated *impedimenta*, which in my opinion was a reasonable mistake.

It's worth mentioning that the corpus was written by a single author with a distinct British style. Even though the differences between American English and British English weren't tested or discussed in this paper, as the percentage of BP generics out of all the plural nouns was distinctly low, I would assume that this trend applies to American English novels as well.

References

Carlson, G. 1988. Truth-Conditions of Generic Sentences: Two Contrasting Views, University of Rochester.

Carlson, G., and Pelletier, F. J. (eds.) 1995. The Generic Book. Chicago: University of Chicago Press.

Carlson, G., 2010. Generics and concepts. In F. J. Pelletier (ed.) Kinds, Things and Stuff. In the New Directions in Cognitive Science series, Oxford. 16-35.

Cohen, A. 2002. Genericity. Linguistische Berichte, 10:59-89.

Dahl, Ö. 1995. The marking of the episodic/generic distinction in tense-aspect systems. In Carlson and Pelletier (1995) 412–425.