Measuring Uncertainty in Social Science Texts

Sarah B. Bouchat
bouchat@northwestern.edu

September 2018

Not For Circulation or Citation

Abstract: Previous work on the quantitative analysis of text emphasizes validity and error of classification schemes for text, usually via comparison to human coding approaches (e.g., Benoit, Laver, and Mikhaylov 2009 and 2012, and Grimmer and Stewart 2013). These works quantify uncertainty with respect to our inferences about information contained in text, but do not engage directly with the uncertainty characterized by the words and tone of the corpora themselves. Using techniques for identifying Type II and higher fuzzy quantifiers (e.g., "about a half," or "rather high") in natural language, this paper proposes a method for measuring uncertainty in text with broad applicability to the social sciences. The paper discusses several applications of the proposed approach to problems and techniques in social and political science.

For their helpful comments on previous drafts, I would like to thank Brandon Stewart and Will Lowe. All remaining errors are my own.

1 Introduction

1.1 Problem & Motivation

Uncertainty estimation is not only core to generating social science inferences, it also fundamentally informs research design. While uncertainty estimates may be commonplace in quantitative analyses, characterizing the uncertainty present in qualitative research or qualitative sources of data, such as text, would provide substantial new opportunities for innovative research. Rather than only evaluating uncertainty at the level of individual variables, social science texts contain and express uncertainty about broad theoretical constructs and relationships between phenomena of interest. For example, in quantitative published works, authors may express uncertainty textually in order to explain anomalous outcomes or data limitations that threaten inference. A reliable way to quantify and evaluate this additional uncertainty would aid in the production of future research. For qualitative research or qualitative sources of data such as interviews, leveraging verbally or textually expressed uncertainty could provide an external mechanism for weighting and validating evidence.

Despite the promise of utilizing textual data for uncertainty analyses, uncertainty estimation in text analysis presents a particular problem because of the high degree of dimensionality in the data and because the structure of otherwise useful assumptions constrains analogues to sample-and-population definitions of uncertainty.

Defining the problem of uncertainty estimation for text data requires distinguishing between (at least) two levels: uncertainty within text and uncertainty about text. The first type of uncertainty—within-text—is the expressed uncertainty of the author(s) of a given document, via word choice, semantic construction, or omission. The second type of uncertainty—about text—can itself reflect the researcher's uncertainty about the measured or detected level of within-text uncertainty (i.e., how accurately automated

methods detect the latent uncertainty of an author and disaggregate it from other explicit or implicit orientations), or can encompass uncertainty about the text at the level of the research construct (i.e., a level of confidence in the inferences drawn regarding the research question given the available data, and/or how well the selected data capture the concepts that the research agenda requires).

Previous work on the quantitative analysis of text emphasizes validity and error of classification schemes for text, usually via comparison to human coding approaches (e.g., Benoit, Laver, and Mikhaylov 2009; Grimmer and Stewart 2013). These works present possible ways to quantify about-text uncertainty, but do not directly measure or evaluate within-text uncertainty. While within-text uncertainty evaluation at first suggests a separate approach, emphasizing word selection/omission as well as arrangement, within-text and about-text uncertainty are not so easily disaggregated in theory. That is, current validation approaches rely on a notion that a crisp, "true" state of a document exists and that only measurement error or uncertainty interferes with inference. An author's latent disposition or intent, however, qualifies the concept that quantitative research belatedly attempts to measure. The extent to which this poses a problem for the validity of social science inference depends in part on the chosen mode of analysis—failing to detect underlying uncertainty or hedging may not interfere as much with recovering the topic(s) in a document as it would with estimating other quantities of interest, such as polarity or ideology.

By contrast, a separate stream of literature focused on international security and intelligence has dedicated significant effort to identifying terms that correlate with uncertainty ("words of estimative probability") in order to quantify risk assessments, but has not adequately addressed the ways in which authors express uncertainty intentionally, unintentionally, and via omission.

Using techniques for identifying fuzzy quantifiers (particularly Type II and higher)

in natural language, this paper proposes approaches for measuring uncertainty in text with broad applicability to the social sciences. In particular, quantifying uncertainty in textual sources can aid methodological research evaluating open-ended survey questions or interview responses; facilitate the weighting of documents and the evaluation of veracity in archival research; leverage data in documents with unreliable or unknown narrators more effectively; and even lay the groundwork for the formalization of these insights from published literature in the form of fully specified Bayesian priors.

While fuzzy logic and fuzzy quantifiers have had a place in computational linguistic theory for many years, applications of the concept both witin the field and in cross-disciplinary endeavors are limited. Even the most straightforward approach of simply detecting and counting uses of fuzzy quantifiers within text runs headlong into a debate about whether, when, and which stopwords to optimally exclude prior to analysis (Manning, Raghavan, and Schütze 2008; Saif et al. 2014; Denny and Spirling 2018). Excluding words such as "all," "any," "few," "about," "some," or "much" might seem prudent to focus on words of relatively greater value in distinguishing across topics or documents, but these and other terms of fuzzy quantification serve as a fundamental basis for expressions of uncertainty about the subject matter within documents.

1.2 Overview for the Paper

The next section reviews literature relevant to uncertainty estimation from divergent perspectives in computational linguistics, computer science, mathematics, philosophy, and social science, to propose that fuzzy quantification offers significant promise in efforts to characterize uncertainty for social science text data. The following section then proposes two broad approaches to implementing uncertainty measurement using fuzzy quantifiers, and the conclusion provides a discussion of directions for additional research and refinement.

2 Literature: Uncertainty in Text

Uncertainty detection and characterization in natural language has drawn attention in computational linguistics, information fusion, and math and computer sciences, where providing theoretical underpinnings and methods have similar aims but disparate formulations, as well as social and security studies scholars, for whom the immediate practical significance is more acute. As such, there remains a gulf between the formal theorization of uncertainty within fuzzy logic frameworks dating to the 1970s and 80s, and successful applications, which are few in number and often end at the stage of detection/classification (e.g., Li, Gao, and Shavlik 2014; Jean et al. 2016; Conde-Clemente et al. 2017).

2.1 Subjectivity & Security Studies

Within NLP, subjectivity analysis—whether at the sentence or word level—has sought to identify "private states" (emotions or beliefs) from textual cues (Breck and Cardie 2014, 3). Information fusion and computational linguistics are home to many ongoing efforts to generate a definitive typology of uncertainty and ambiguity that would undergird detection efforts (Rubin, Liddy, and Kando 2006). Akkaya, Wiebe, and Mihalcea (2009) offer insights into the problem of subjective word sense disambiguation, which plagues efforts to accomplish accurate automated subjectivity analysis. Identifying words that reflect sentiments or opinions, rather than words that "cue" subjectivity but are intended objectively, is challenging precisely because subjectivity is a "private state"—much like uncertainty, it is a latent condition that can only be partially observed or measured via language (2). While the authors introduce computational approaches to correctly identifying subjectivity, the conceptual overlap between subjectivity and uncertainty is limited and the attempt to classify language into binary usage categories (subjective/objective) is very limiting in the context of attempting to characterize either the type or extent of

uncertainty reflected in language. For example, the authors discuss two differing uses of the word "catch," where "What's the **catch**?" reflects a *subjective* use of the term meant to indicate a "drawback," whereas "He sold his **catch** at the market" is instead *objective* because it represents a quantity. For the purposes of uncertainty estimation, however, the indefiniteness of the quantity "a catch" is of greater importance than its objective sense (Akkaya, Wiebe, and Mihalcea 2009, 2). The subjectivity lexicon emerging from this research agenda, therefore, offers a very incomplete inventory of the types of words that lend themselves to uncertainty quantification.

Dragos (2013, 4) provides a useful overview distinguishing among types of "intrinsic" uncertainty:

(1) **Ambiguity** arises naturally in language because it can never truly capture all observations and internal experiences, but in particular emerges from polysemy or incompleteness.

(2) **Vagueness & Precision** are contrasting concepts that may reflect intentional or unintentional shifts toward or away from exactness. For example, "they have several cats" is a vaguer construction of "they have 4 cats," where the choice to use "several" in the place of a discrete quantifier may reflect a lack of knowledge about the precise number or may be an intentional choice to shift importance (e.g., if the emphasis is that they have cats rather than dogs).

Rather than focusing on strict linguistic ambiguity—the kind of uncertainty that arises from gaps between the signifier and the signified, as with polysemy and homonymy—or reality-limited referential ambiguities—such as when color descriptions are inherently biased, unclear, or culturally dependent (e.g., green-blue versus blue-green versus turquoise, or proximate versus obviative referents)—treating what Dragos calls the "vagueness" form of uncertainty as a *knowledge domain* contained within language implies several clear modes of approach, beginning with lexical definition and detection,

and proceeding through quantification (Auger and Roy 2008, 1862–1863, 1865–1866). Druzdzel (1989) catalogued 178 verb phrases and modifiers indicating uncertainty (e.g., "it could be," "it seems," "impossible"), a pursuit informed by the interest of the security community in quantifying uncertain levels in natural language, stemming from the famous 1973 Rand report by Sherman Kent.

Probability (percent)	Verbal Equivalent
100	It is certain that ...
85-99	It is almost certain that ...
60-84	It is probable that ...
40-59	The chances are about even that ...
15-39	It is probable that ... not ...
1-14	It is almost certain that ... not ...
0	It is impossible that ...

Figure 1: A chart from the 1973 Kent report, as depicted in Auger and Roy (2008, 1866)

The security studies lineage of quantifying qualitatively expressed uncertainty in text arises from the immmediate, significant implications for precisely characterizing experts' assessments of the likelihood of events has immediate, significant implications (Miller et al. 2013). Recognizing the need to specify and examine uncertainty across multiple dimensions and applications, Thomson et al. (2005) developed a framework that disaggregates nine distinct types of uncertainty: accuracy, precision, completeness, consistency, lineage, currency, credibility, subjectivity, and interrelatedness (Auger and Roy 2008, 1861). As with Dragos' distinction, not all of these types are necessary or feasible for quantifying uncertainty in social science texts and applications. For example, assessing the accuracy or completeness of analyses in social science articles or in open-ended survey responses may be beyond the scope of automation, and evaluating the credibility of a source may require information not contained within the source text itself.

Beyond theoretical distinctions among types of uncertainty, Druzdzel's survey pro-

vides evidence of a human preference to express uncertainty qualitatively rather than numerically, perhaps as an instinctual desire to avoid overly certain statements. Practically speaking, this evidence underscores the significance of accurate methods for uncertainty detection and measurement—however defined—in natural language processing.

2.2 Fuzzy Logic & Fuzzy Linguistics

Mathematics and formal logic, in turn, do not lack for ways to conceptualize and reason with vagueness and uncertainty (e.g., Wygralak 1998; Roos 1990), but correlating these theoretical constructs with expressions in natural language presents significant challenges (Barwise and Cooper 1981; Zadeh 2005). In particular, common statements such as "most people" or more abstract phrases such as "there are only a finite number of stars" cannot conform to standard first-order logical conditions such as $\forall x(...x...)$ (Barwise and Cooper 1981, 160). This shortcoming indicates precisely why fuzzy logic and possibility theory have offered more promising avenues of theoretical exploration.

Fuzzy logic, while maligned by proponents of traditional logic schema across disciplines, directly leverages and incorporates the ambiguity inherent in language into formal reasoning. Unlike in classical logic, where propositions are evaluated as either true or false, fuzzy logic propositions have a "truthfulness" value that is a real number on the interval $[0, 1]$ (Atanassov and Gargov 1998, 40). While the use of fuzzy logic is often at odds with formalized probabilistic representations (e.g., Lindley 1987), its aims and application bear clarification:

> Fuzzy logic is not fuzzy. Basically, fuzzy logic is a precise logic of imprecision and approximate reasoning. More specifically, fuzzy logic may be viewed as an attempt at formalization/mechanization of two remarkable human capabilities. First, the capability to converse, reason and make rational decisions in an environment of imprecision, uncertainty, incompleteness of information, conflicting information, partiality of truth and partiality of possibility—in short, in an environment of imperfect information. And second, the capability to perform a wide variety of physical and mental tasks without any measurements and any computations. (Zadeh 2008, 2753)

Natural language epitomizes the "imperfect information" that fuzzy logic and fuzzy sets attempt to contend with. Fuzzy linguistic representations have presented significant challenges to implementation, however, with fragmented efforts either emphasizing a single test application that does not generalize, or focusing on questions beyond uncertainty per se (e.g., Cabrerizo et al. (2017) utilizing fuzzy linguistics to evaluate "consensus" in decision-making). In particular, though, the evaluation of fuzzy quantifiers has been one critical point of entry for bringing fuzzy logic theory to linguistic applications, originating in the the study of generalized quantifiers in language (Barwise and Cooper 1981). Fuzzy quantifiers, in general, are imprecise descriptors often used to construct propositions for fuzzy logic. Terms such as "most," "many," or "few" provide a general sense of the quantity in question, but only within certain bounds and sometimes with reference to a vaguely defined population (analogous to the cardinality of a fuzzy set[2]). The degree of imprecision and the reference category, in turn, characterize the type of fuzzy quantifier, or the degree of fuzziness (versus crispness). Specifically, Zadeh (1983, 150) defines a fuzzy quantifier as "a fuzzy number which provides a fuzzy characterization of the absolute or relative cardinality of one or more fuzzy or nonfuzzy sets."[3]

Distinguishing among types of quantifiers, furthermore, is both necessary for evaluating not just *whether* statements are uncertain but the *degree* to which they are, and is also an ongoing effort. Zadeh (1983) posits at least three types of fuzzy quantifiers, but broadly, type or degree varies according to the fuzziness of the predicate and the quantification in the phrase (Liu and Kerre 1998, 2):

Type I: Cardinal number extensions of two-value logic (e.g., "All buildings are

[2]The cardinality of a fuzzy set can be a real or a fuzzy number, but broadly speaking, represents the number of elements that are members of a set S. Conceptually, the cardinality of a fuzzy set is complicated, as it can itself be fuzzy and therefore require characterization only in relation to other fuzzy sets (Dhar 2013).

[3]That is, not all fuzzy numbers are fuzzy quantifiers, and whether a fuzzy number qualifies as a fuzzy quantifier is sometimes unclear.

solid," "Greater than half of the surface of the southern hemisphere is water")

Type II: Quantifiers of fuzzy sets (e.g., "Some children are tall")

Type III: Fuzzy quantifiers of crisp sets/ratios of fuzzy quantifiers (e.g., "Almost all mammals indigenous to North America are eutherian,")[4]

Type IV: Fuzzy quantifiers of possibility distributions (e.g., "Few politicians remain popular long")

Dìaz-Hermida, Bugarìn, and Barro (2003) goes further in examining the bounds of these quantifiers (semi-fuzzy quantifiers, ultra-fuzzy quantifiers, and the process of quantifier fuzzification), while others have posited a multitude of possible ways to detect and measure these types of quantifiers (e.g., Szmidt and Kacprzyk 2001; Zhai and Mendel 2011; Chen, Song, and Heo 2017; Ramos-Soto and Pereira-Farina 2017). For the purposes of this paper, "Type II" fuzzy quantifiers will serve as the primary point of entry for developing a method of uncertainty measurement and evaluation. Many critical concepts and research agendas in the social sciences can be thought of in the context of fuzzy sets: scholars often draw theoretical battle lines around definitions that constitute set members (e.g., what is a democracy versus an autocracy, what types or magnitude of violence qualify as war, etc.). Statements on the basis of empirical evaluations of these theoretically defined concepts, then, directly map onto fuzzy quantification. Likewise, fuzzy or "soft" quantification is more flexible to changing population sizes than traditional metrics, and as such may facilitate easier model comparison (Farnadi et al. 2016, 61).

2.3 Possibility Theory

More broadly, how fuzzy quantifiers are expressed and evaluated hinges on the underlying possibility distribution that defines the fuzzy number. That is, informally, overlap

[4]Likelihood ratios also qualify as a type of fuzzy quantifier.

exists in our conception of what constitutes "most" and "more than half," but how much these categories overlap depends on the ease with which we believe a number that is $> 50\%$ belongs to the set we conceptually define as "most." That is, while both 52% and 80% are contained in the set "more than half," colloquially we may believe that the descriptor "most" is more easily attributable to 80% than 52%.

Possibility theory provides a framework through which to adjudicate precisely these distinctions, occupying conceptual space between probability and fuzzy set theories. As Khoury, Karray, and Kamel (2008) articulates:

> A probability of 0 means a certainty that the event will never occur, while a probability of 1 is a certainty that the event will occur. Fuzzy set theory, on the other hand, deals with the membership of an event in a set. A membership value of 0 means that the event does not belong at all in the set, while a membership value of 1 means that the event epitomizes the set. Finally, the possibility theory expresses the ease with which an event can occur, or belong to a set. A possibility of 0 means that an event cannot occur, while a possibility of 1 means that the event is completely allowed to occur, and values between 0 and 1 represent events that are restricted but not impossible. (1531)

While applications of possibility theory to text analysis are few and far between, its direct relationship to the ambiguity of language make it conceptually compelling. For example, Khoury, Karray, and Kamel (2008) track the domain representation of particular words in a corpus via a possibilistic domain classifier. That is, with language stripped into subject-verb-object triplets, they instantiate the domain information of language in the form of possibility distributions. As an example application, they use triplets gathered from descriptions of learning objects in the SchoolNet corpus to conduct classification, calculating a possibility ($\pi(d_t)$) that indicates the ease with which a given test learning objective can belong to the domain d_t (1540). Because their test is performed with a winner-takes-all approach, they only evaluate the statistic with respect to the most-possible domain. Beyond utilizing the subject-verb-object construction, therefore, the practical utility of this approach relative to topic modeling or other alter-

natives is less clear. Conceptually, however, using possibility theory and fuzzy sets to both directly and indirectly characterize uncertainty is promising. Indirectly, possibility theory can generate flexible bounds or constraints on the subject matter of interest or important words or topics and can thereby represent uncertainty. Khoury et al. offer the example phrase "this room is cold," where "cold" represents a fuzzy set that bounds our expectations about the likely temperature of the room. If the temperature were 15°C, we could state that the constraint is satisfied to the degree 0.8. Framed in possibility theory terms, this would indicate that there is a *possibility* of 0.8 that the room temperature is 15°C given that it was described as "cold;" that is, 0.8 suggests the "ease" of labeling a room at that temperature as "cold" (Khoury, Karray, and Kamel 2008, 1532).

More directly, possibility theory could serve to construct a lexicon of uncertainty terms or to refine existing lexica related to subjectivity, such that terms are classified not as representing uncertainty per se, but rather the degree to which they indicate uncertainty (e.g., modifiers such as "never" are not compatible with the set of terms that characterize uncertainty whereas terms like "about" may be compatible with that set at 0.5, where polysemy accounts for other uses and membership of the term in other sets). Both practically and theoretically, the continuous character fuzzy quantifiers and possibility theory is more attractive than in the discrete labeling common for subjectivity analysis.

In this paper, I contend that fuzzy logic and fuzzy linguistic principles are not only themselves potential measures of uncertainty in text, but also can be used as measurement *tools* to assess uncertainty within and across texts. The following section presents empirical approaches to applying fuzzy logic and fuzzy quantifiers in the context of providing uncertainty estimates for textual data.

3 Methods for Uncertainty Estimation with Fuzzy Quantifiers

This section discusses two broad categories of approaches to utilizing fuzzy quantifiers in uncertainty estimation for social science. The first category focuses on dictionary-based approaches that are familiar to subjectivity analysis, while the second set proposes a more expansive adoption of fuzzy logic principles to assess uncertainty in textual data. Applications to other types of social science texts are discussed following the initial analysis.

Both broad sets of approaches to using fuzzy quantifiers ultimately require detecting fuzzy quantifiers in text, and utilizing fuzzy quantifiers as a measurement tool imposes extra weight and significance on the accuracy of measuring the quantifiers themselves. While more work is likely necessary to define and refine the set of fuzzy quantifiers most useful and relevant to estimating uncertainty of various types, this section proceeds from a practical and inclusive perspective that will be amenable to later changes in specification. No global dictionary of fuzzy quantifiers in English currently exists, so I take several steps to generate an inclusive set set of fuzzy quantifiers:

(1) Generate a list of terms used in the original Zadeh (1983) article examining fuzzy quantifiers

(2) Add items from Liu and Kerre (1998), which reviews and explains Zadeh (1983) with additional examples

(3) Add empirical examples from corpora used for testing (BioScope corpus and Farnadi et al. (2016)) and empirical applications (Conde-Clemente et al. (2017)):

   (a) Test existing list from items (1) and (2) to recover items from corpora

   (b) Hand-review Bioscope corpus sample for additional examples

The comprehensive list of fuzzy quantifiers used in this section is provided accord-

ing to source in the Appendix.

3.1 Dictionary-Based Approaches

3.1.1 Fuzzy Quantifiers & Subjectivity Terms

As indicated previously, the literature that attempts to empirically evaluate uncertainty almost exclusively focuses on dictionary-based methods. These efforts largely center on subjectivity analyses that identify and classify non-objective words and phrases; rarely are fuzzy quantifiers included, let alone central to the analysis. At a more theoretical level, the subjectivity analysis approaches implemented to date do not emphasize quantifying subjectivity at the document or corpus level—this quantification would require a weighting or aggregation scheme from the word- and sentence-level detection procedure. This limits the utility of current approaches for characterizing how uncertain a given article or author might be, above and beyond simply answering the question of whether they use sentences containing uncertainty terms.

A simple first step toward bridging subjectivity analysis and fuzzy quantifier approaches is to assess whether and to what extent fuzzy quantifier terms are distinct from and improve upon subjectivity terms. For convenience, this step reproduces, in part, the 2010 shared task from CoNLL, which required the detection of hedge words (Farkas et al. 2010), and data from Vincze (2014), which implements uncertainty detection across several corpora. Vincze uses the Szeged Uncertainty corpus ("Szeged Uncertainty Corpus" 2010), which is composed of the subcorpora BioScope 2.0, FactBank 2.0, and WikiWeasel 2.0 (details in Table 3.1.1). These datasets are useful primarily because they are pre-labeled for subjectivity, but attempting to "predict" that pre-labelled value with only fuzzy quantifiers is not feasible on theoretical as well as empirical grounds. On a theoretical level, as indicated by the aforementioned literature, the type of uncertainty indicated by fuzzy quantification is often conceptually distinct from what is coded as "subjective."

One dataset is primarily occupied with automatically detecting the *type* of subjectivity at the word level, therefore requiring several aggregation assumptions in order to recover a document-level metric of "uncertainty," as well as assumptions about which of these types is most compatible with the quantifiable "uncertainty" that fuzzy quantifiers embody. On a practical level, furthermore, the rates of fuzzy quantifier usage are relatively low across all of the pre-labeled corpora, making detection and classification particularly challenging.

Corpus	Description	Labels	Size
Bioscope	Radiology reports, full biology papers, and Genia corpus abstracts	Negation, hedge words, and hedge word scope	20,924 sentences
FactBank	Newswire and broadcast data	Events coded into 4 "factuality" types by source and point of view	3,123 sentences
WikiWeasel	Paragraphs from the full Wikipedia dump	"Weasel" cues	20,745 sentences

Figure 2 illustrates the overall rate of fuzzy quantifier usage by corpus, scaled by the total number of words. While variation exists, the tight clustering around an average rate of 0 underscores the challenge of conducting a validation task comparing fuzzy quantifiers with subjectivity words. Likewise, at the document level, the average number of fuzzy quantifiers per sentence is relatively low, with similarly low variance, as shown in Figure 3. An examination of the association between fuzzy quantifiers and subjectivity words, as shown in Figure 4, is inhibited by the overall low rate of fuzzy quantifier usage, but additionally suggests that at most, fuzzy quantifiers only capture a single facet of subjectivity, rather than being a completely congruent concept. Likewise, while the narrative structure and expression of uncertainty likely differs by corpus, fuzzy

quantifiers and at least some kinds of subjectivity terms more likely function as substitutes rather than complements when articulating uncertain views. As such, a correlation measure based on co-occurrence at the sentence or document level is unlikely to capture their theoretical relationship.

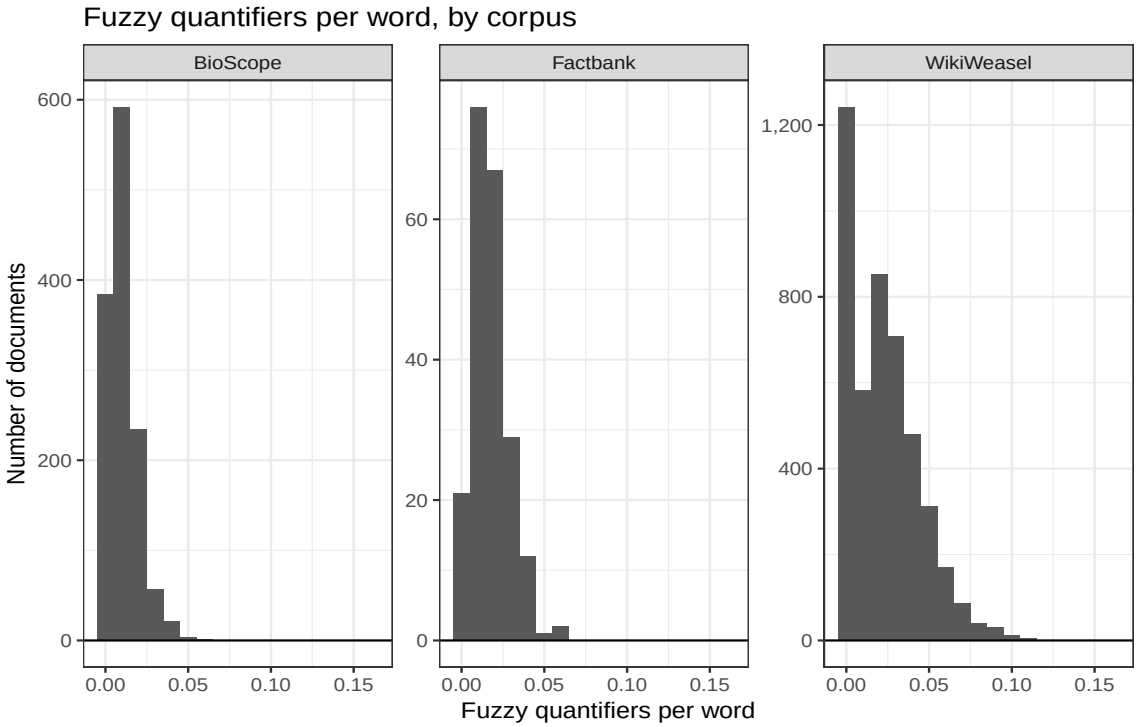

Figure 2: Per-Word Fuzzy Quantifier Rates by Corpus

Fuzzy quantifiers, then, provide an opportunity to refine the measurement of expressed uncertainty apart from the generic detection of "subjectivity" more traditionally in use. Because fuzzy quantifiers modify particular terms or phrases, however, they also require a method of measurement and detection more sensitive to their specificity, rather than exclusively focusing on their existence or nonexistence. Furthermore, different fuzzy quantifier terms serve to express different types and magnitudes of uncertainty, attention to which is required for measuring uncertainty as the underlying concept.

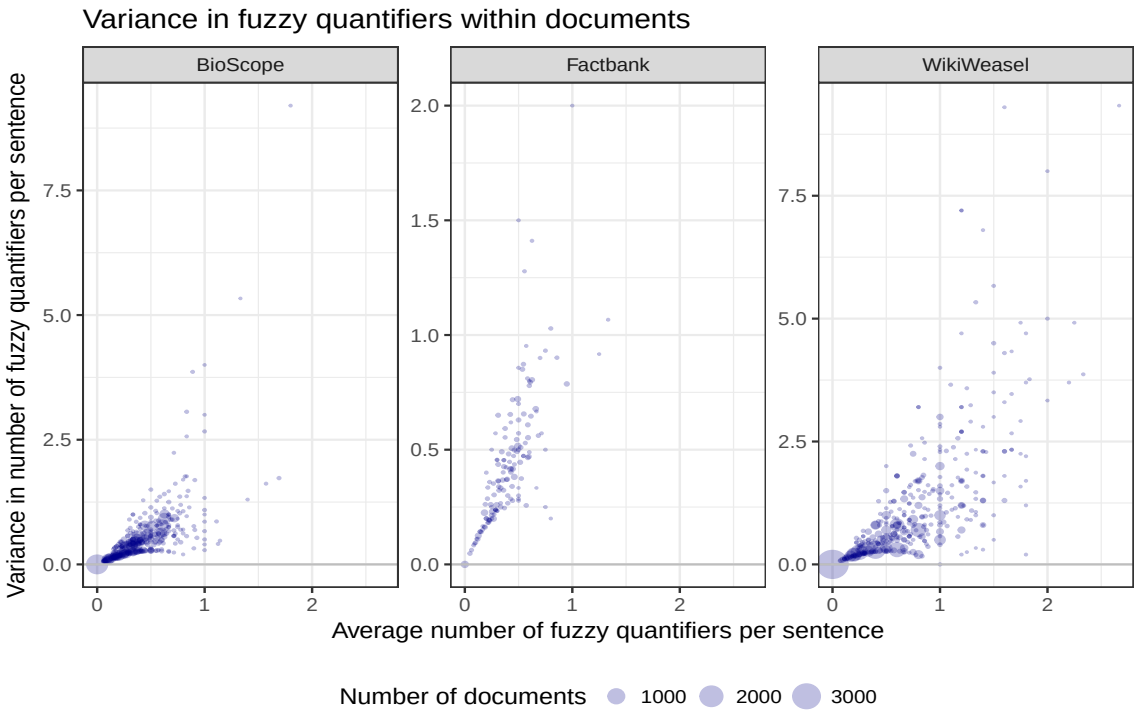

Figure 3: Fuzzy Quantifier Variation (Avg. Num. Per Sentence) by Document

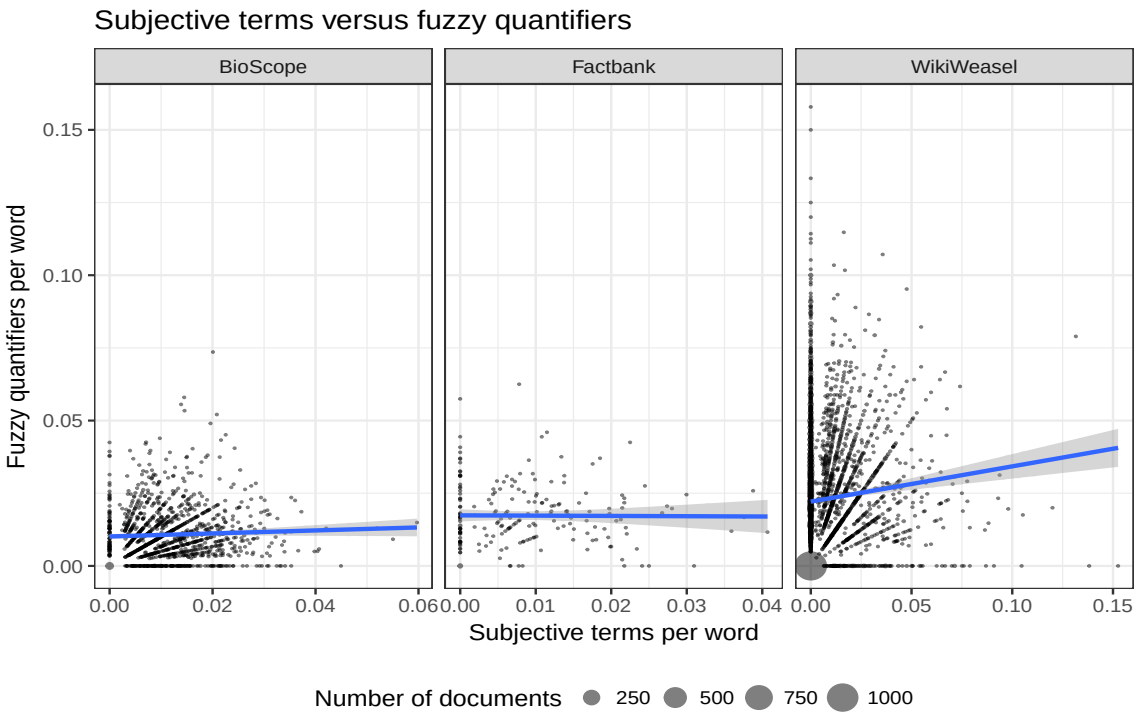

Figure 4: Correlation between Fuzzy Quantifiers and Subjectivity Terms

3.1.2 Example: Fuzzy Quantifiers as Keyword Modifiers

To better illustrate the ways in which authors use fuzzy quantifiers to express uncertainty, this section presents evidence from a corpus of newspaper articles across a variety of tagged subjects. The corpus comprises 20,667 English-language articles (5,953,382 words) from the Burmese newspaper, *The Irrawaddy*, spanning from 29 March 2012 through 18 August 2018. The articles are subjected to minimal preprocessing, removing only special characters/symbols, stopwords exclusive of fuzzy quantifiers (see Appendix), and stemming.[5] The corpus is tokenized as five-grams as an initial approximation of the likely window in which fuzzy quantifiers are used to modify key subject-related words.

To identify the "keywords" that fuzzy quantifiers modify, articles are grouped according to their subject tags. These tags appear as text at the end of articles, preceded by "Topic." Because these subject tags can be extremely specific (e.g., "Kofi Annan," "timber smuggling," "mass graves," "cassette tapes"), the set of subject tags used in this analysis is limited to the top 20 most commonly occurring. This does allow for some article overlap across topics (e.g., approximately 40 articles share the tags "Conflict" and "Rohingya"). From these top 20 subject tags, the top 20 most commonly occurring words in each are selected and used to subset the five-grams to those featuring at least one key word (11,696,686).

Grouping articles within their sections and tagged subjects serves multiple purposes. First, journalists and editors at the newspaper likely have a consistent beat, meaning that articles on the same subject may have patterns of fuzzy quantifier usage related to authorship that are distinct from subject matter. Second, some subject matter (e.g., politics, conflict) is more likely to reflect uncertainty than others (e.g., lifestyle articles).

[5]Removal of stopwords is conducted to maintain consistency between the identification of "top keywords" by subject tag and words identified in five-grams.

Evaluating fuzzy quantifier usage across these dimensions allows for assessing these assumptions while also providing structure to temporal comparisons. Finally, to the extent that fuzzy quantifiers truly expressing uncertainty are theorized to modify words directly related to critical subject matter, restricting the set of keywords by tagged subject should provide a more accurate measure of fuzzy quantifier usage and uncertainty.

Within each of these top 20 subject tags, for five-grams containing at least one relevant key word, I evaluate co-occurrence of fuzzy quantifier terms according to their varying types (see Appendix). Results are presented below as a series of comparisons. Figure 5 through Figure 11 present the ratio of occurrences of each fuzzy quantifier term relative to the other fuzzy quantifier terms of its type ($FQ_i : FQ_{-i}$), where the assumption is that authors choose between several possible quantifiers to express their level of uncertainty about a given keyword or subject conditioning on the appropriate type as well as magnitude. That is, if the author is expressing uncertainty about a proportion, they choose from among fuzzy quantifiers of proportions and not from the full set, including fuzzy quantifiers for quantities such as time or relationships.

From this set of comparisons, a few patterns of interest emerge to consider for future assessments of uncertainty. For the set of time quantifiers, "often" appears between 2 and 10 times more often than all other fuzzy time quantifiers; and "no" appears as commonly or more often than other proportion quantifiers. In the case of "no" and "between," these sharp contrasts to other fuzzy quantifiers of the same time indicate that greater specificity in the usage of the term is necessary for measuring uncertainty directly (that is, rather than also including instances of "between" that do not imply an approximate range, for example).

In addition to indicating the need for greater refinement of terms and their usage, these results suggest that a metric of uncertainty requires indicating degree even within the type of uncertainty. Where numeric quantifiers are concerned, "many" and "some"

are significantly more common than other possible quantifiers in relative terms while not being absolutely as likely as all other numeric quantifiers combined. That is, if the numeric set of fuzzy quantifiers were further delineated into those that imply larger or smaller amounts or greater or lesser degrees of uncertainty, "many" and "some" are likely to be more prevalent and preferred measures. Greater specificity among these fuzzy quantifiers could likewise facilitate controlled comparisons across types; that is, evaluating the conditions under which, for example, an author might specify "many" rather than "most" where "most" likely implies greater certainty about the quantity of interest.[6]

[6]In this case, an author may have uncertainty about the numerator *or* the denominator of a proportion, perhaps indicating a greater likelihood of trading off for numeric quantifiers versus proportion quantifiers.

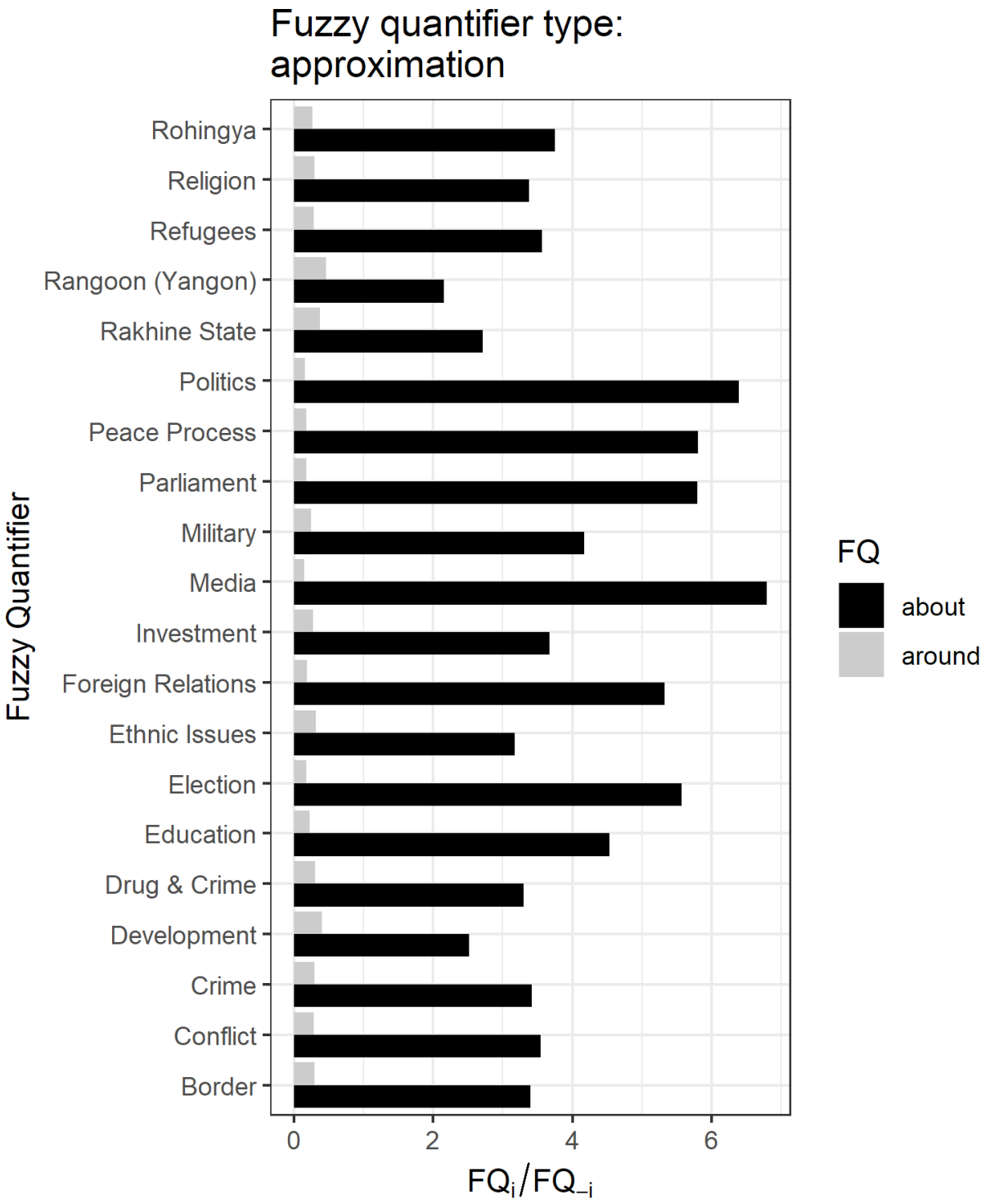

Figure 5: Ratio of "approximation" fuzzy quantifiers across fivegrams, by topic

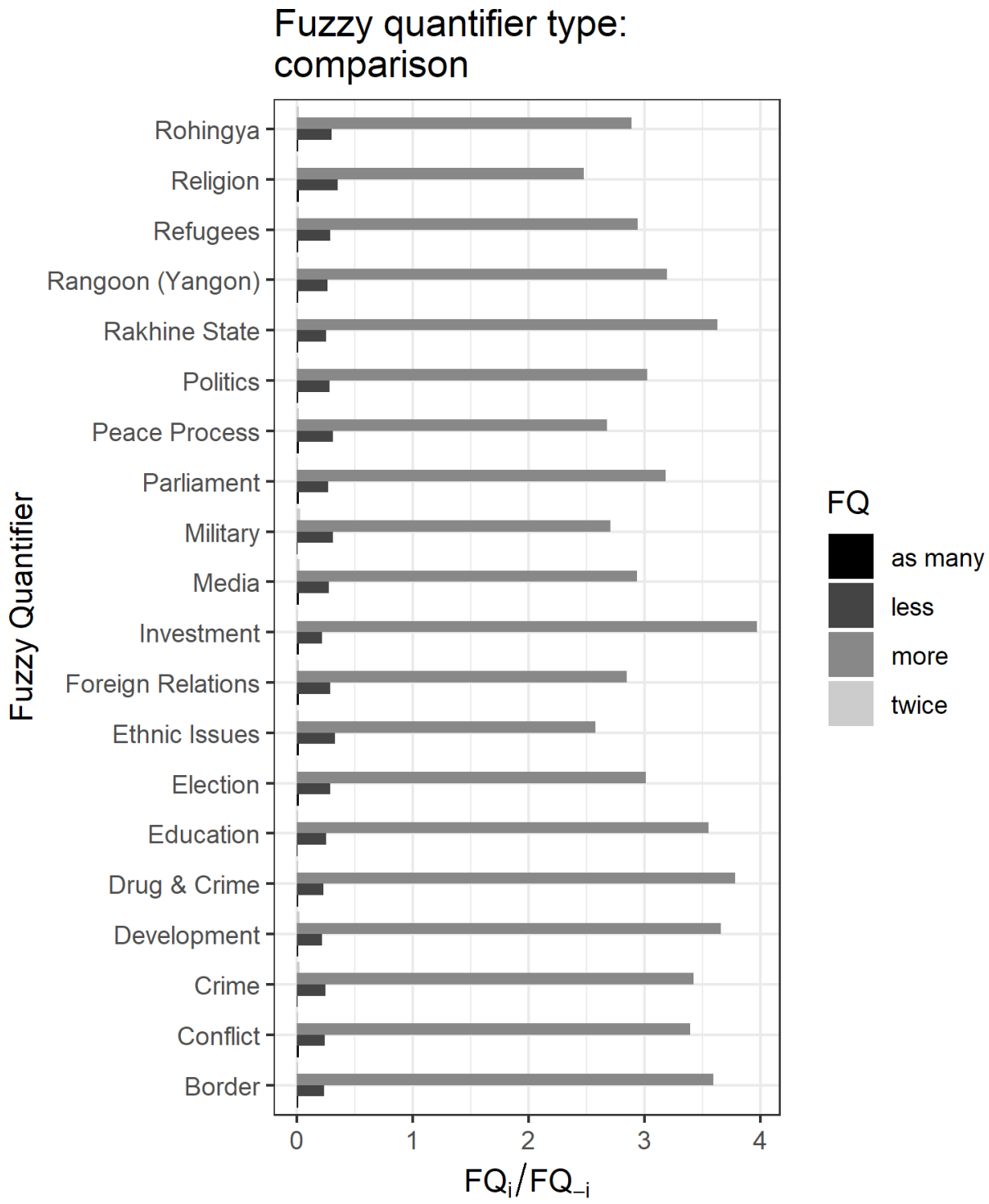

Figure 6: Ratio of "comparison" fuzzy quantifiers across fivegrams, by topic

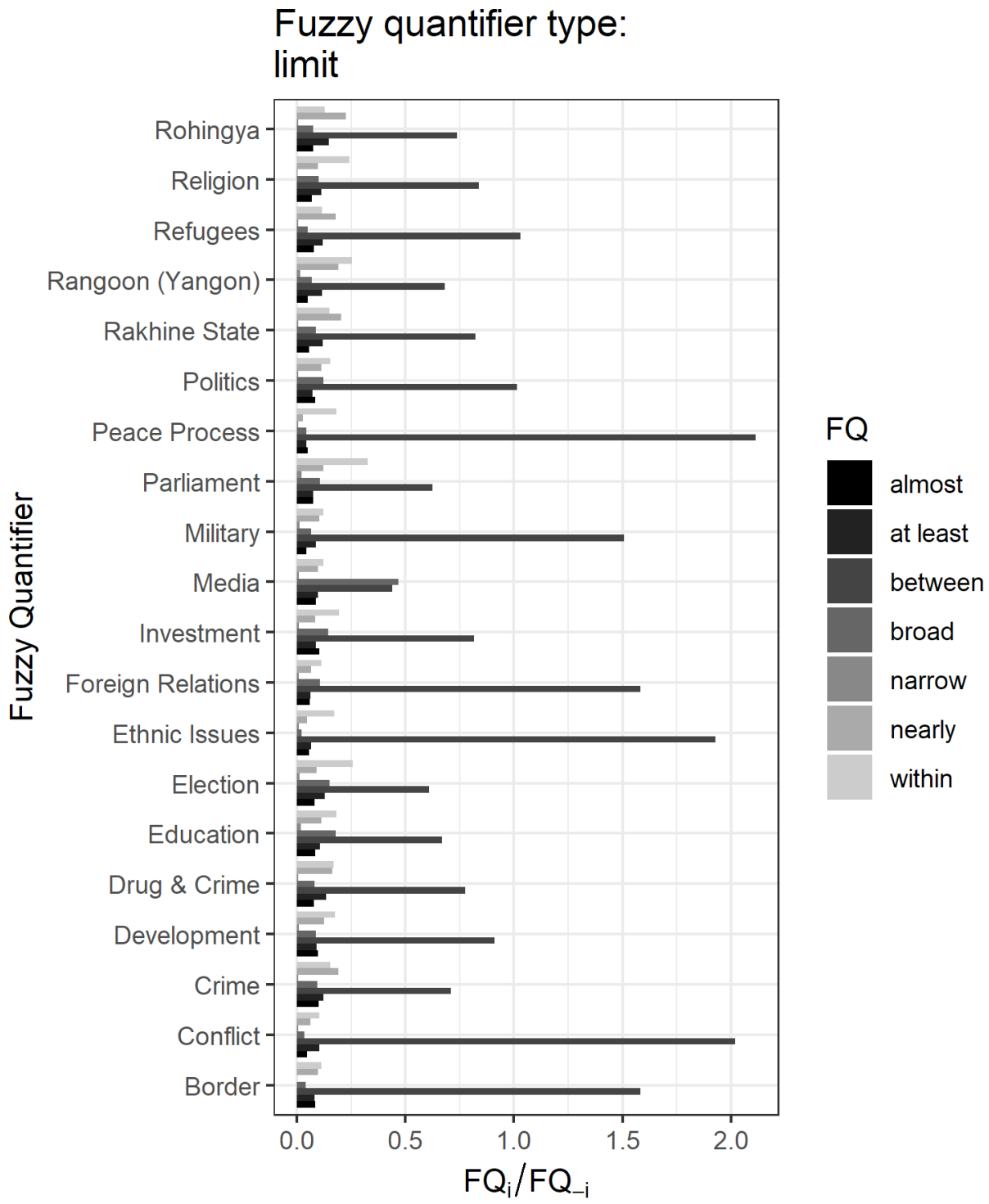

Figure 7: Ratio of "limit" fuzzy quantifiers across fivegrams, by topic

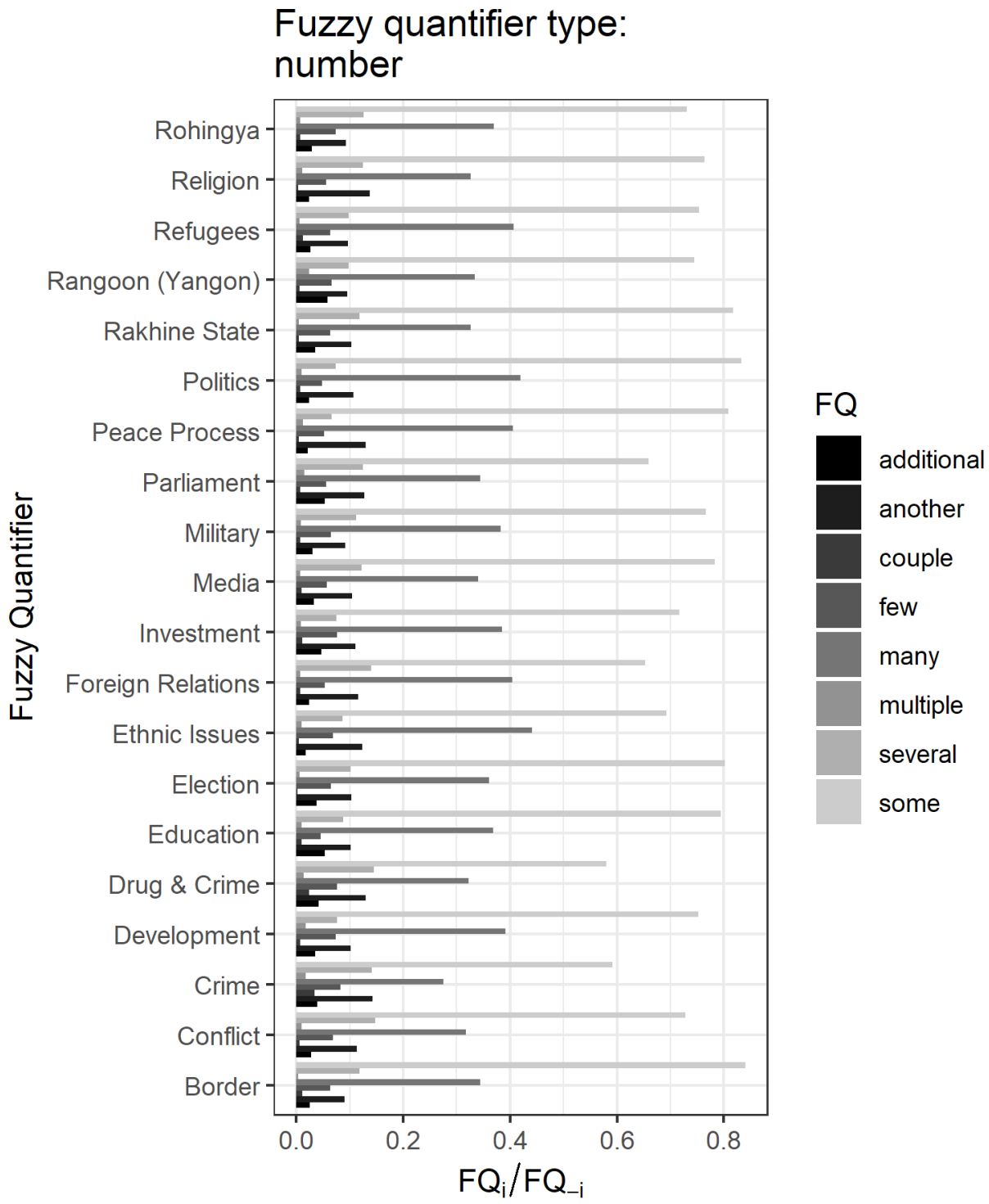

Figure 8: Ratio of "number" fuzzy quantifiers across fivegrams, by topic

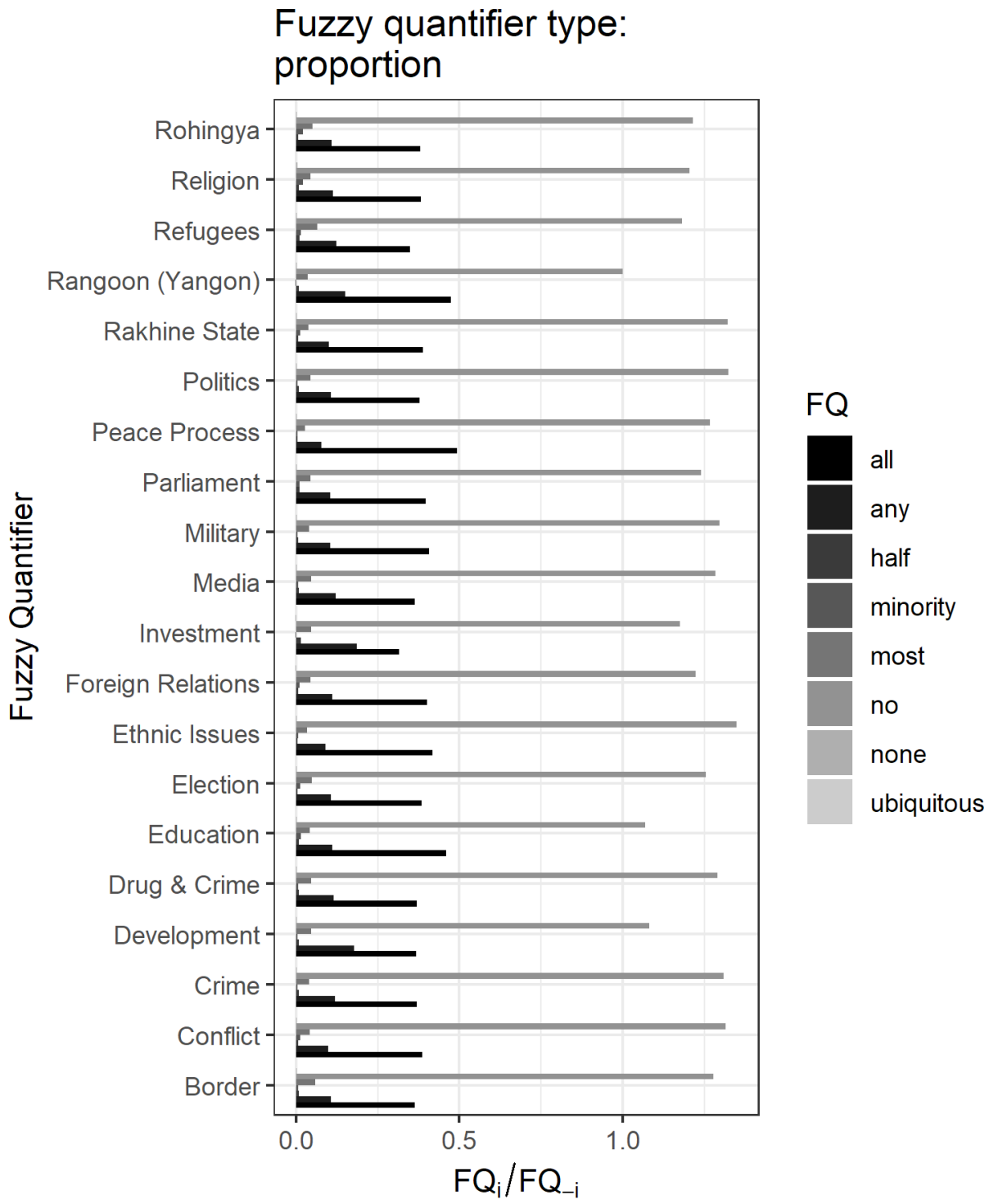

Figure 9: Ratio of "proportion" fuzzy quantifiers across fivegrams, by topic

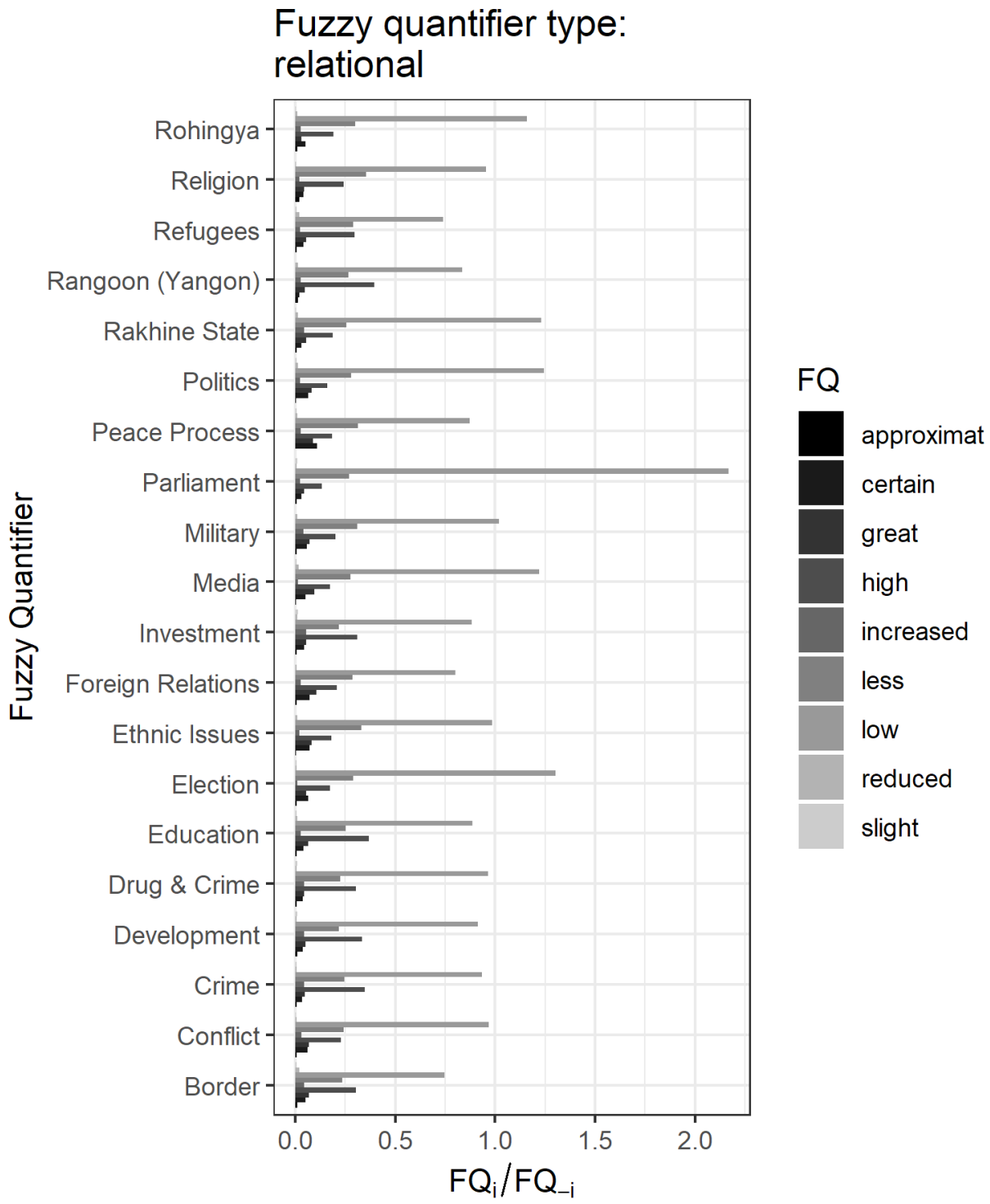

Figure 10: Ratio of "relational" fuzzy quantifiers across fivegrams, by topic

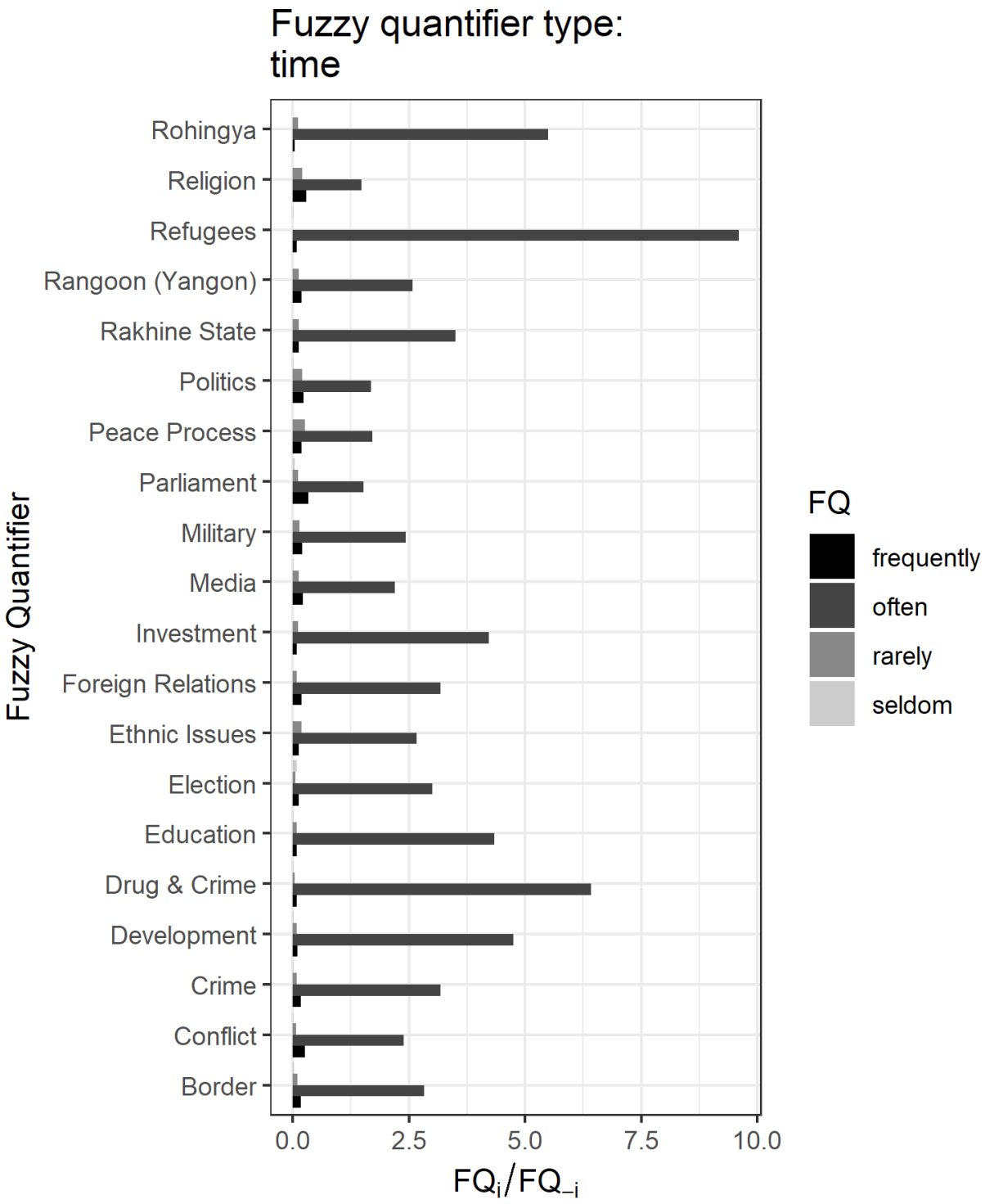

Figure 11: Ratio of "time" fuzzy quantifiers across fivegrams, by topic

Figure 12 through Figure 18 in turn present ratios of five-grams containing a given fuzzy quantifier (FQ_i) to those not containing a fuzzy quantifier that still contain at least one key word for the subject tag. While the comparisons within fuzzy quantifier type demonstrate significant variation in the use of fuzzy quantifiers across subjects and within specific kinds of uncertainty, the comparisons to five-grams that contain no fuzzy quantifiers broadly illustrate the rarity of fuzzy quantifier usage as before. All values, irrespective of fuzzy quantifier type, remain close to 0, indicating the five-grams without expressed uncertainty are much more likely than those with it. This comports with expectations for the expression of information in news media generally: news intends to be informative rather than speculative and should prefer exact statements to uncertain ones, particularly when accounting for a temporal dimension (that is, it is preferable to offer no statement and later an exact one than to provide early uncertain statements, particularly in print journalism).

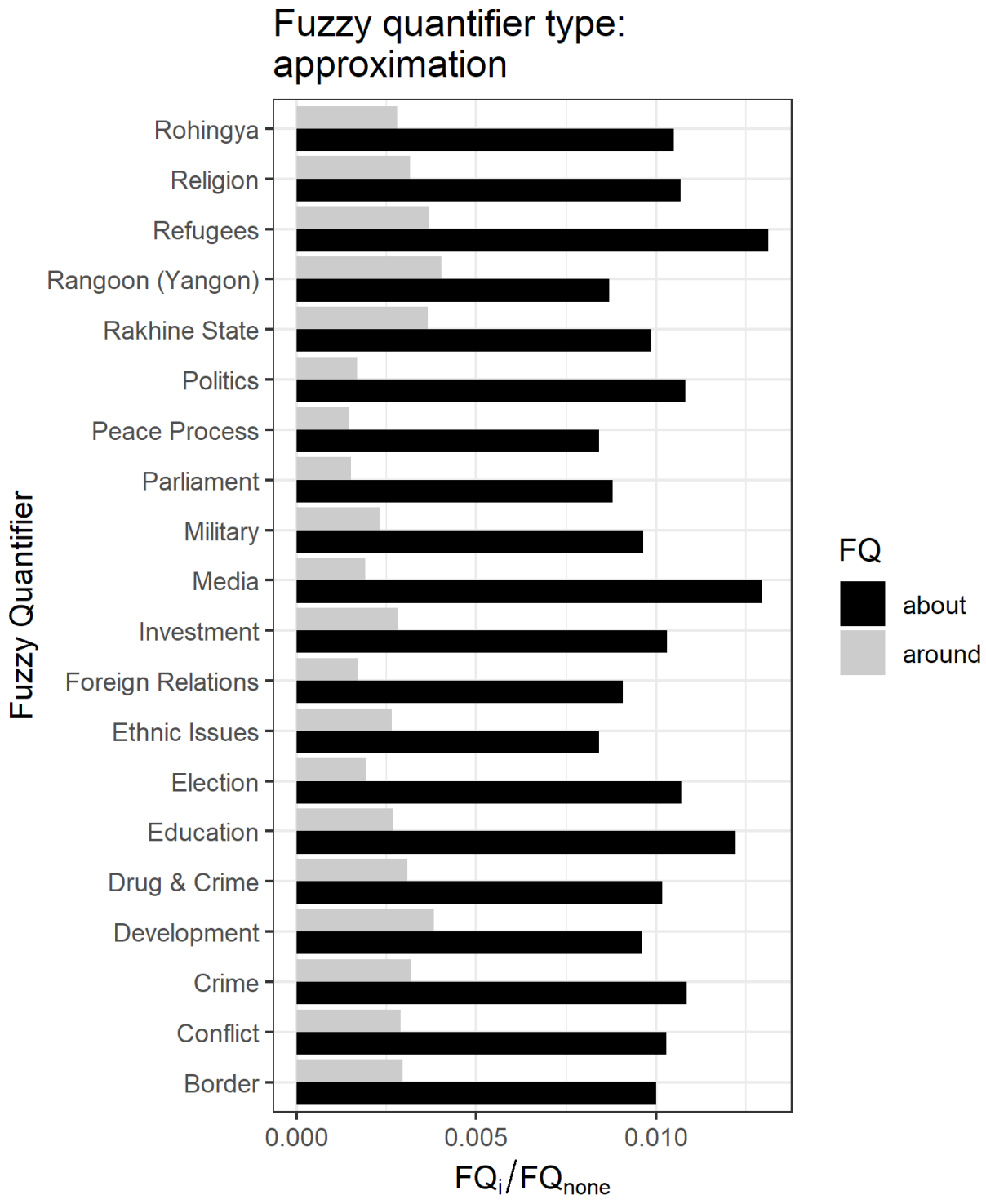

Figure 12: Ratio of "approximation" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

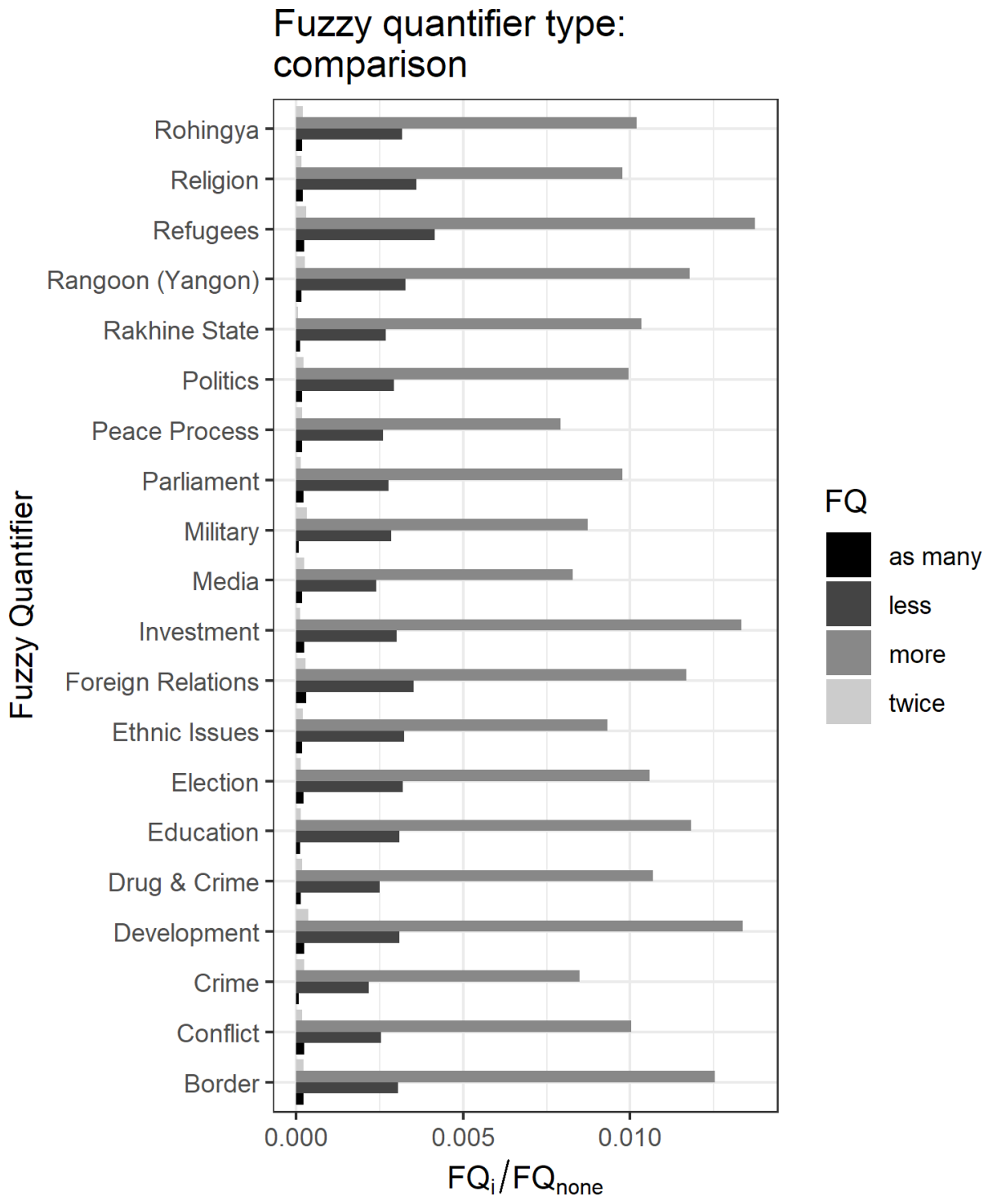

Figure 13: Ratio of "comparison" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

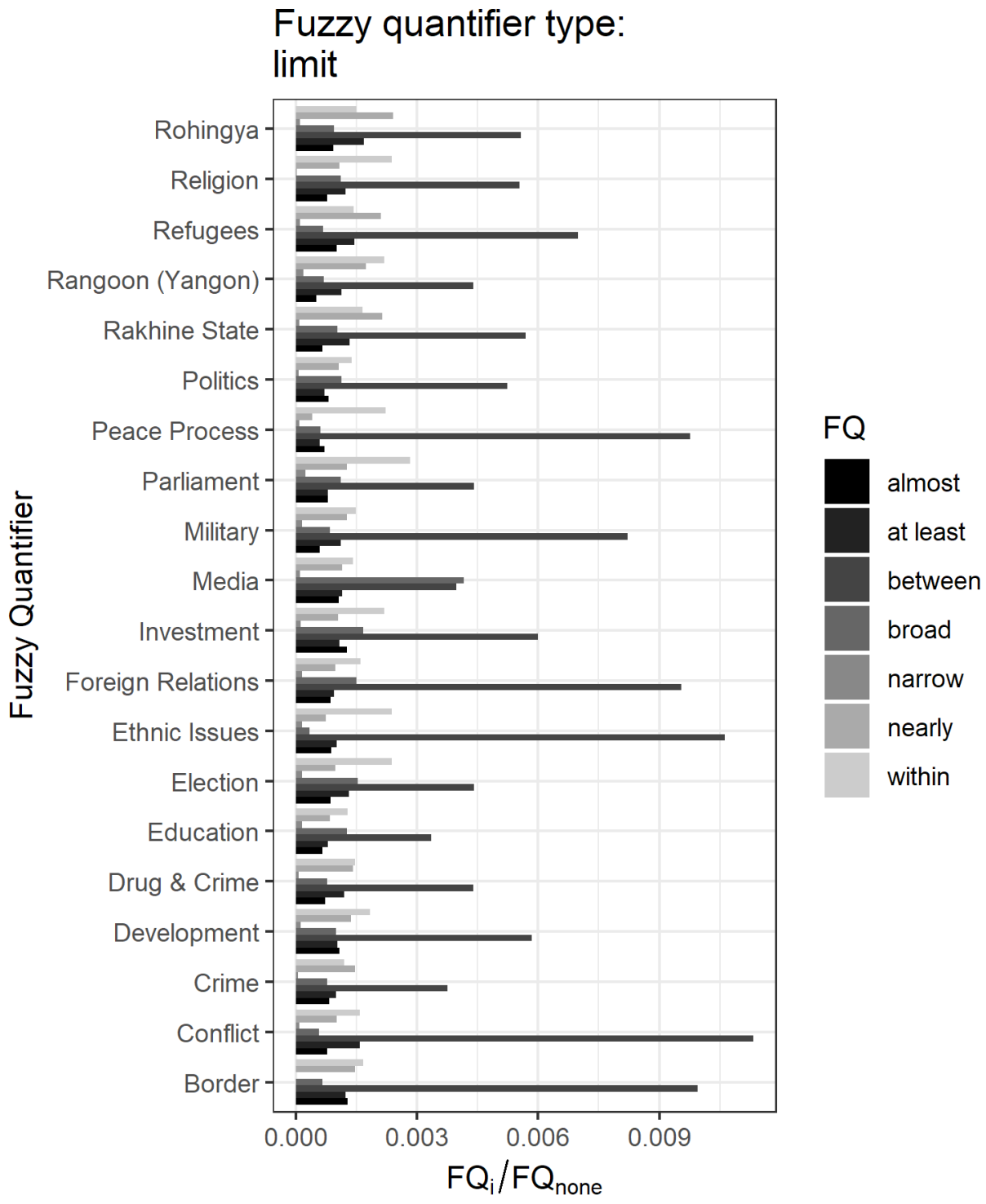

Figure 14: Ratio of "limit" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

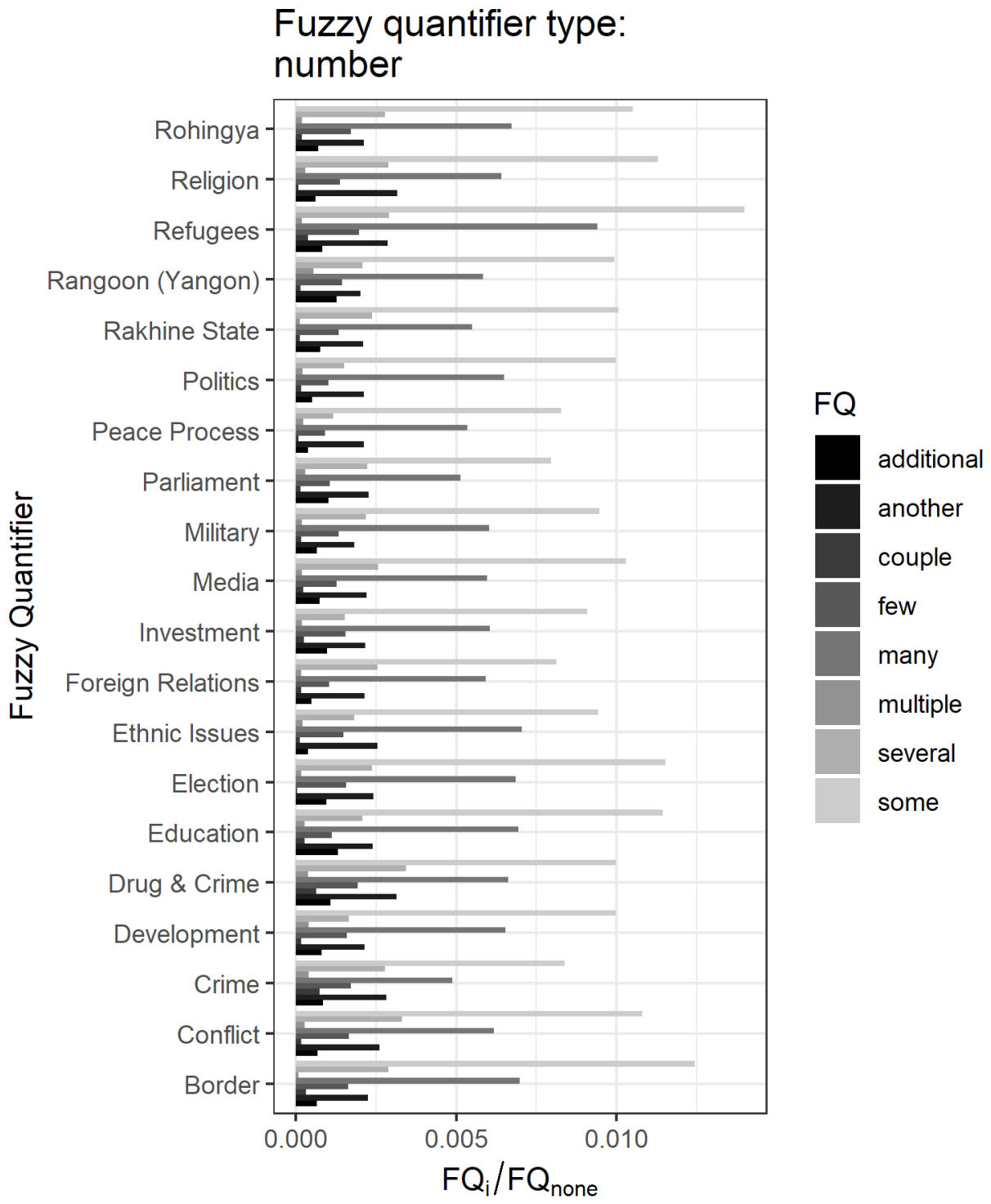

Figure 15: Ratio of "number" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

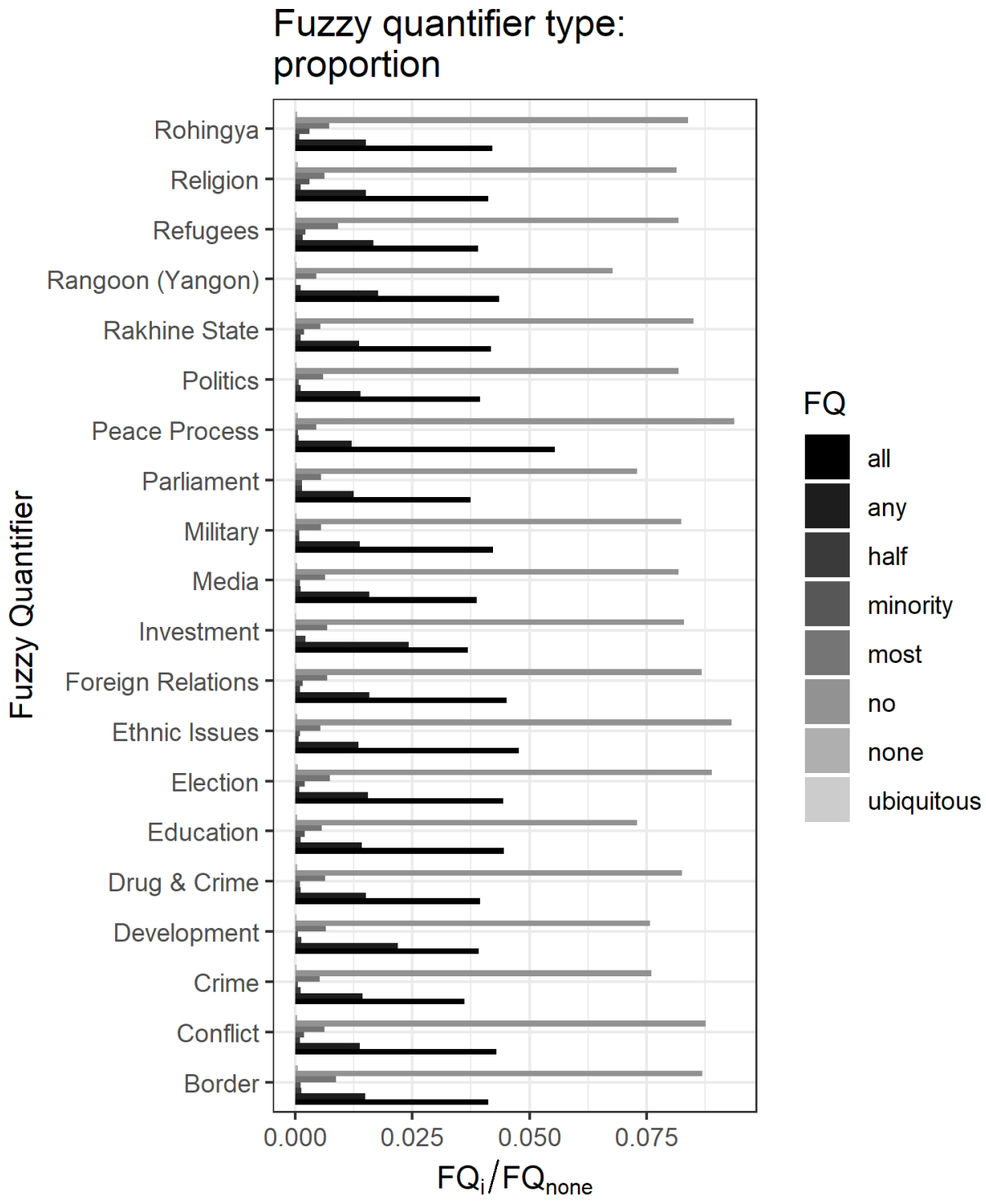

Figure 16: Ratio of "proportion" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

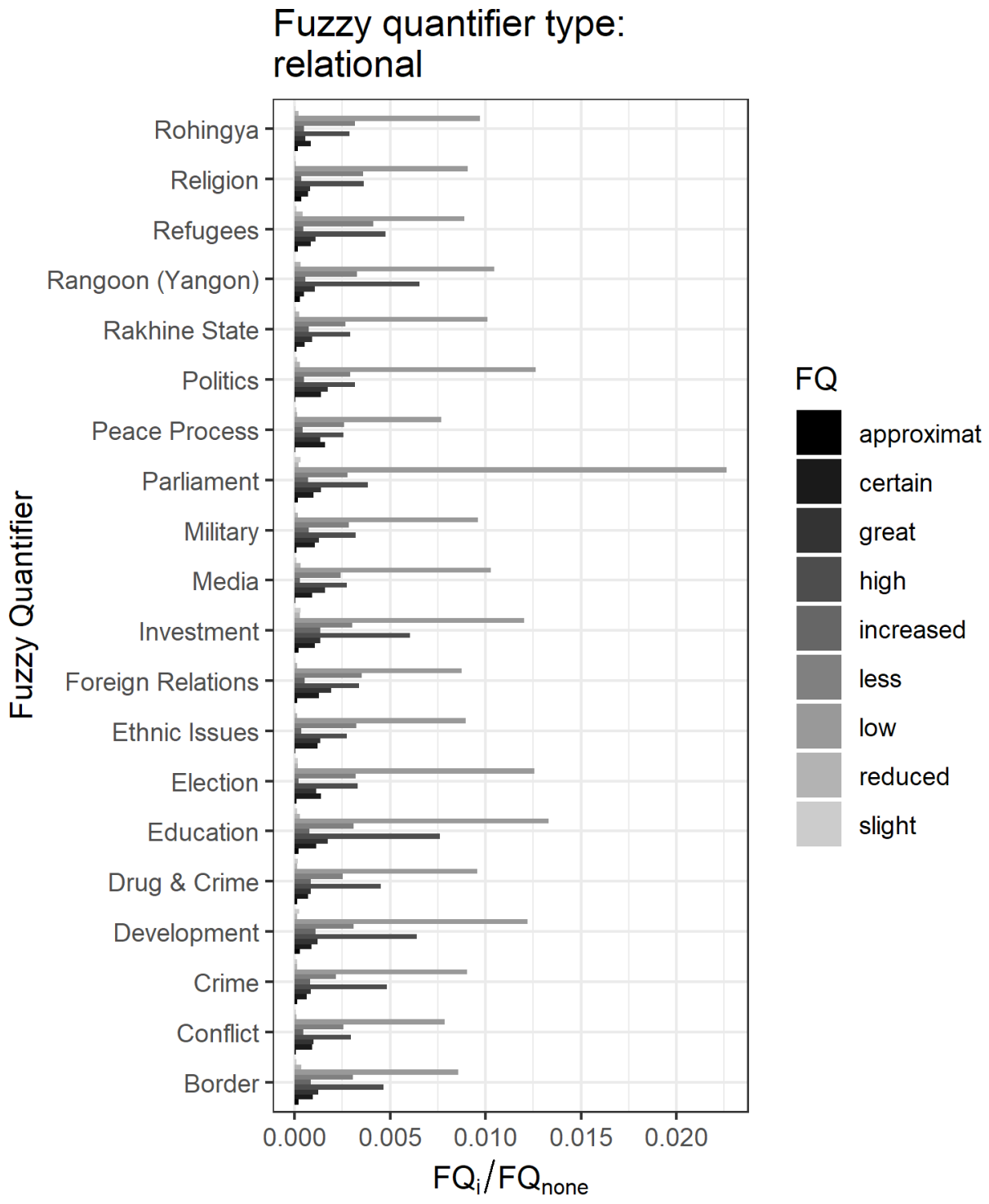

Figure 17: Ratio of "relational" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

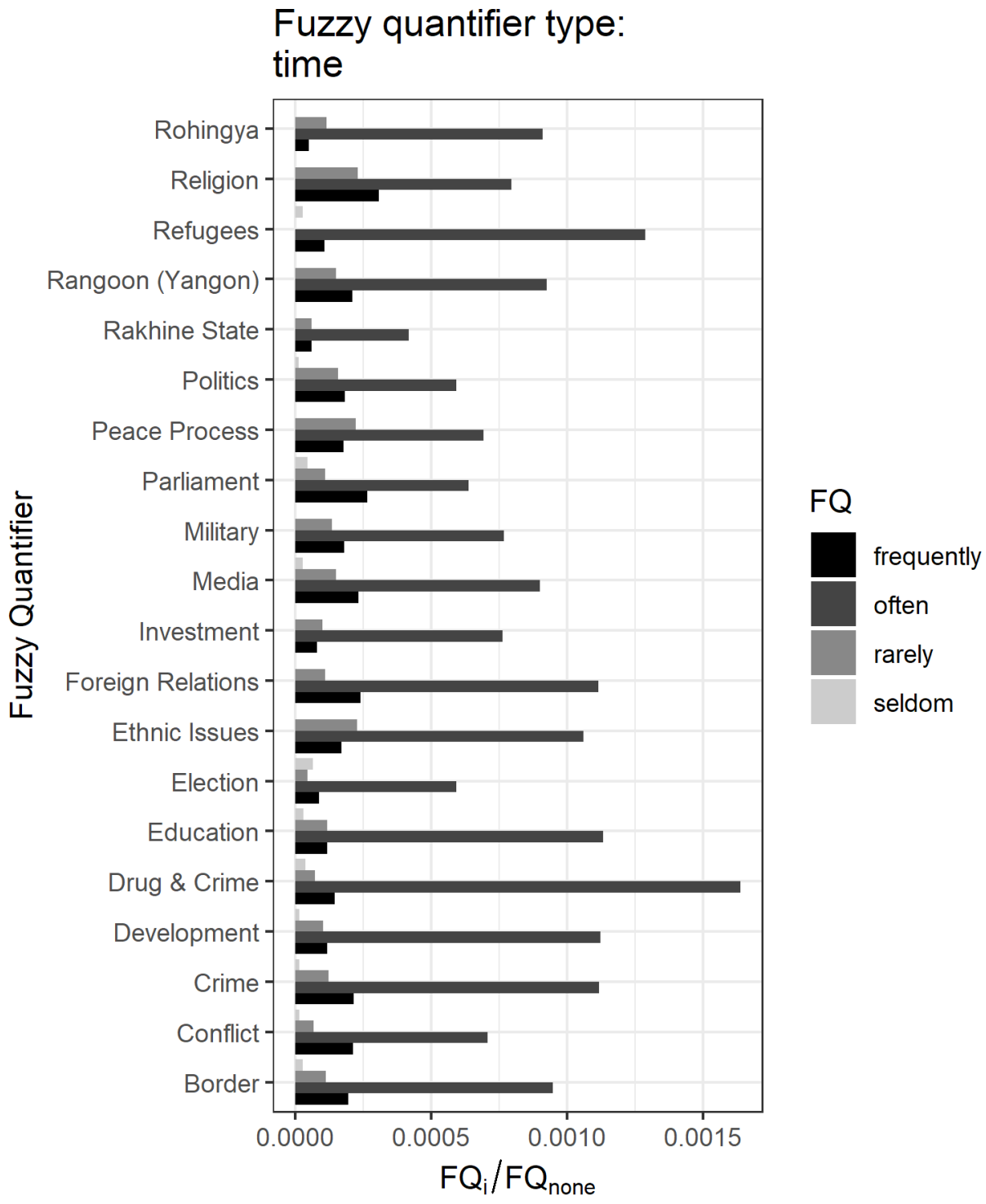

Figure 18: Ratio of "time" fuzzy quantifiers across fivegrams to fivegrams with no fuzzy quantifiers, by topic

The assumption concerning the broadly informative aim of news media likely does not hold for other types of text relevant to social science research, or indeed to other types of "news" media. *The Irrawaddy* is broadly assumed to be engaged in equilibrium behavior from a Bayesian semantics perspective because it is one of few "free" media sources in Myanmar, with some foreign funding and at least some foreign audience. The same could perhaps not be said of government-controlled news sources in authoritarian regimes, where deception or obfuscation are acceptable characteristics of discourse.

The sparsity of fuzzy quantifier terms in these texts serves as a cautionary lesson in directly applying simple counts, or perhaps even emphasizing only co-occurrence of fuzzy quantifier terms with words or phrases of interest. Rather, measuring a level or degree of uncertainty given their sparsity requires greater attention to the specificity of the uncertain statement and the choice of fuzzy quantifier relative to other possible modifiers. At a minimum, this requires a more complete dictionary of fuzzy quantifiers as well as non-fuzzy quantifiers.

3.1.3 To Quantify or not to Quantify

As indicated previously, the usage of fuzzy quantifiers does appear to differ significantly across topics, and this topical difference likely also relates to the incentives of news media for "precise" reportage in various areas. That is, the "tolerance" for less precision or more uncertainty likely varies by topic: when reporting on a shooting, news media are expected to provide details such as "32 people wounded" rather than "several people wounded" where possible, but reporting on other topics may not place such a premium on precision. This topical disparity is further illustrated by Figure 19. While the Wikipedia data previously referenced lack the same sort of "subject tags" in use in the articles from *The Irrawaddy*, the average fuzzy quantifier usage in the Wikipedia sample is a reasonable baseline from which to see the variation in fuzzy quantifier usage according to topic. To the extent that fuzzy quantifiers serve to provide less specific,

more speculative assessments in text, this variation comports with expectations: articles about border issues, conflict, and refugees appear more "fuzzy" than, e.g., articles about investment or crime. Broadly speaking, "precision" correlates more strongly to a concept of "certainty" than "imprecision" does to "uncertainty," but where incentives exist to provide highly specific content and imprecise or fuzzy terms are nevertheless used, this should capture uncertainty in the reporting. The incentives faced by news media in this case are further supported by market competition among news outlets. For example, when reporting on how many of the soccer team members were safely freed from captivity in a Thai cave in the summer of 2018, news media repeatedly updated despite initial uncertain reports until they achieved greater precision (e.g., "as many as 3 boys have been freed," eventually became "5 boys are now safe").

What this analysis lacks, in part, is a robust theory of the incentives and motivations for using fuzzier or less precise quantifiers relative to other possibilities. A simple comparison of precise versus fuzzy quantifiers illustrates the disparity in their usage across topics as incentives shift. To structure this comparison, I identify words expressing "precise" quantification as well as numeric values in all articles in each topic set. For "precise" quantification, I include words indicating units of measure (e.g., "minute," "hour," "kilogram," "mile") as well as both word- and numeric expressions of number values ("fourteen" as well as 14, and "million," "billion," etc.). A full list of these "precise" quantifier terms is available in full in the Appendix.[7] Figure 20 indicates that overall rates of quantifier or quantifying term usage are fairly similar, which suggests both a natural rate of usage for quantifying terms and also that some incentive to provide quantification exists within news media. For example, the rate of precise quantifying term usage in articles pertaining to "investment" is nearly twice that of the fuzzy quantifier usage, reflecting the nature of the subject matter but also the premium on accurate and

[7]Several of the units of measure included her are obscure or archaic, but are included for completeness. In later analyses focused on ratios of occurrence, these terms will be removed as a robustness check.

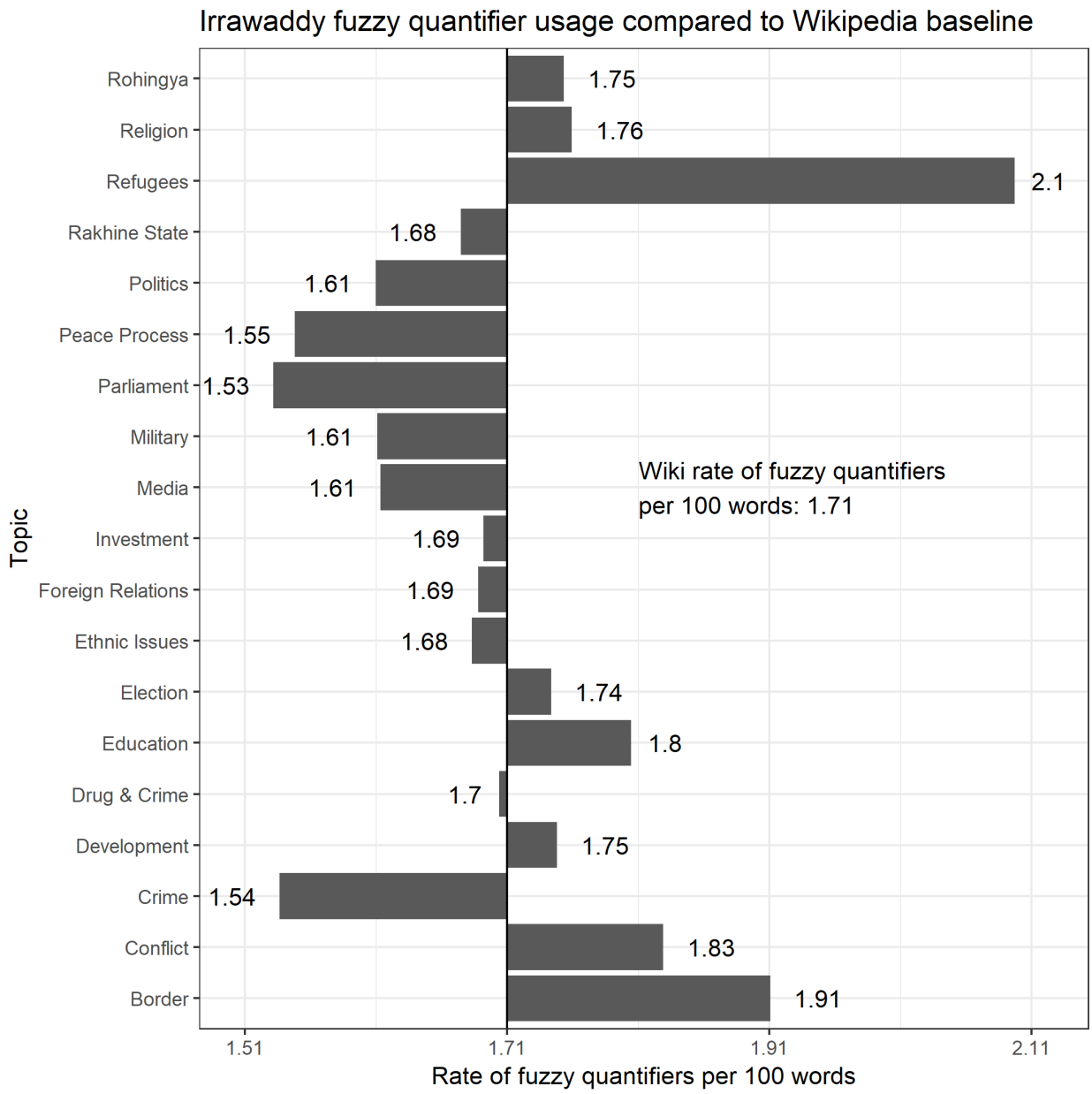

Figure 19: Fuzzy Quantifier Rate per 100 words vs. Wikipedia Baseline, by topic

specific reporting in economic news (e.g., "1.5%" is much more common and acceptable than "a few percentage points"), in comparison to, for example, coverage of conflict, where less specific reporting ("hundreds of deaths," "several thousand troops") may be more tolerable.

These results are preliminary and do not capture the immediate tradeoffs between fuzzy and precise quantifiers and then among fuzzy quantifiers. That is, these figures present aggregate comparisons rather than directly evaluating where opportunities existed to use specific quantification ("80 percent") and fuzzy quantification prevailed ("most"). Accurately specifying these tradeoffs requires a more comprehensive theory of the semantic rules governing quantifier usage (some of which is captured in generalized quantifier theory, particularly in relation to the overall amount of imprecise quantification that can exist in language before an equilibrium of understanding is compromised), but also a theory of the "political economy" of quantification and uncertainty. In particular, the incentives that drive specific versus fuzzy quantification are likely to differ across the type and form of text under examination. While in the news media, incentives to provide more specific quantification abound, a goal in social science publications, for example, may be to make the most general statement possible given the competing specificity of the data. For example, "most democracies are peaceful" is arguably preferable to "60% of democracies are peaceful" conditional on data since the project of social science *is* to generalize.

Articulating the incentive structure of the authors that is likely reflected in the text is also key to understanding tradeoffs *among* fuzzy quantifiers. Extrapolating from an example offered in Lassiter and Goodman (2017) makes this clear:

> Emma is saving 6 cookies for dessert—two for each member of her family. She leaves them on the kitchen counter while she goes to the bank. She calls home while she is waiting in line. Her husband Dan reports to her: "Charlie ate some of the cookies." From the literal meaning of this sentence, there is a very strong inference that (assuming Dan is reliable and well-informed) the

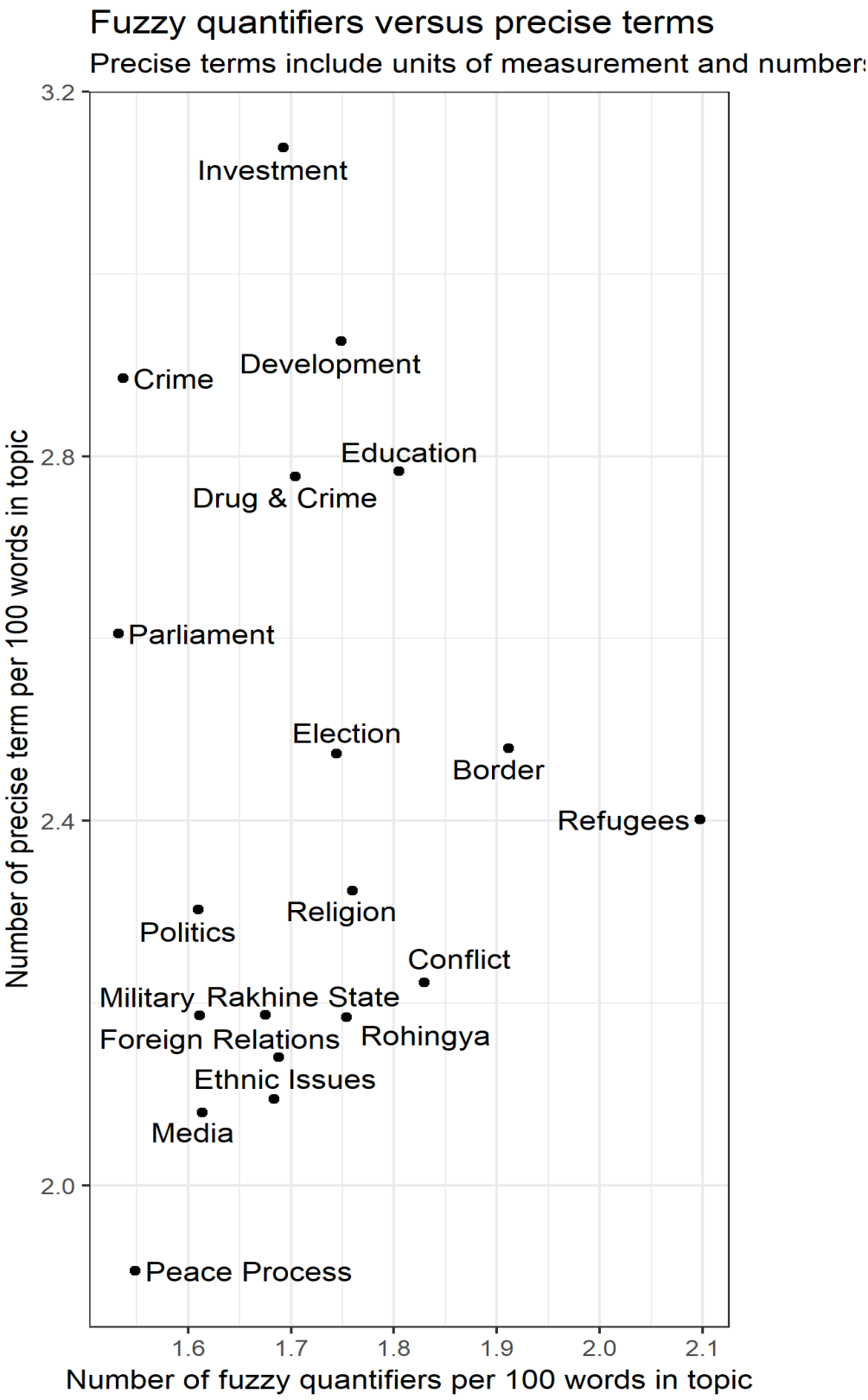

Figure 20: Fuzzy Quantifier Rate per 100 words vs. Precise Term Rate per 100 words, by topic

> number of cookies Charlie ate is greater than zero. Emma will typically acquire more information than this, though: she will also learn that the number he ate is less than six, i.e., he didn't eat *all* of the cookies. Of course this is a defeasible inference—conceivably, Dan is lying or confused. Nevertheless, given what Dan said, Emma can reasonably expect some dessert to be left when she gets home. (Lassiter and Goodman 2017, 3811)

Without abandoning the assumptions that Dan is well-enough-informed and being truthful, applying different possible incentive structures to his revelation suggests differing interpretations, and the possibility of isolating *uncertainty* in particular. For example, Dan need not be lying in order to be using the quantifier "some" to avoid blame or culpability: perhaps Charlie has eaten cookies on Dan's watch before, perhaps this is far more sugar than Charlie (presuming Charlie is a child) is supposed to have in one sitting. In these cases "some" is meant as a hedge ("but not all," "but not that many," "but not too much") where greater specificity would have created interpersonal friction (Emma could interpret "some" to mean 3, but if Dan had said "Charlie ate 5 of the 6 cookies," he may have upset Emma more). Likewise, saying "some" rather than "nearly all" or "most" may be intended to lessen the blow of the news, even when all three quantifiers are plausibly appropriate. Alternatively, Dan was not paying attention previously and does not know the base number of cookies that existed (perhaps he thought Emma made an even dozen), or perhaps there are cookie remnants all over the floor, or perhaps the family also has a dog who was accompanying Charlie on this cookie-eating mission, so the true number of cookies eaten by Charlie is unknown: in this case Dan's use of "some" reflects uncertainty. Locating and specifying the incentives of the author, as well as the opportunities that existed for providing greater specificity, then, is key to determining which uses are truly uncertain rather than vague in incentive-aligned ways.

3.2 Fuzzy Proposition Evaluation

Rather than relying exclusively on dictionary-based approaches and effectively "counting" occurrences of fuzzy quantification, which presents not only theoretical but also

practical challenges, a different approach would more firmly embed uncertainty estimation of text in a fuzzy logic framework prior to leveraging fuzzy quantifiers themselves. Zadeh (1983) offers an example of "test-score semantics," which itself views "everything that relates to natural languages [as] a matter of degree" (152). In a test-score setting, Zadeh proposes generating an *explanatory database frame* (EDF) that contains relational information for pieces of text that can be evaluated via test-scores against propositional statements. These test-scores, then, provide document-level quantities that can measure uncertainty both directly and indirectly. This section applies his hypothetical example to a social science problem for illustrative purposes.

By way of a proposition to evaluate with text, Zadeh offers the hypothetical that "Over the past few years Nick earned far more than most of his close friends," where the relations this requires cataloguing in an EDF include:

> **INCOME**: listing the name of friends, amount they earned, and year
>
> **FRIEND**: listing the name of friends and a value μ that represents the degree to which the person is a friend of Nick
>
> **FEW**: μ representing the degree to which some *number* qualifies as "few"
>
> **MOST**: μ representing the degree to which some *proportion* qualifies as "most"
>
> **FAR MORE**: two income values and μ indicating the extent to which the first income value qualifies as "far more" than the second income value (153).

The information contained in each of these relationships and the elastic constraints imposed by the value(s) of μ facilitate evaluating the "truthfulness" or possibility of the propositional statement. This process very readily applies to social science domains, where this proposition assessment can be thought of metaphorically as hypothesis testing for relationships between variables of theoretical importance. Even so, Zadeh's article is primarily directed toward demonstrating that the concept of fuzzy sets applies

to natural language, and that test score operations and cardinality are likewise *possible*. The article does not, therefore, emphasize ways of measuring or characterizing μ, the parameter that largely determines what inferences one might draw about uncertainty statements within text.

Farnadi et al. (2016) attempts to overcome this shortfall in an application to statistical relational learning (SRL). The paper proposes one family of quantifier mappings that correlate soft (fuzzy) quantifiers with quantitative metrics. Specifically, the paper defines a mapping $\tilde{Q}$ for soft quantifier Q such that:

$$\tilde{Q}_{[\alpha,\beta]}(x) = \begin{cases} 0 & \text{if } x < \alpha; \\ \frac{(x-\alpha)}{\beta-\alpha} & \text{if } \alpha \leqslant x < \beta; \\ 1 & \text{if } x \leqslant \beta. \end{cases}$$

This mapping means that, for example, one could define $\tilde{Q}_{Few} = \tilde{Q}_{[0.1,0.4]}$ (64). This formulation is guided primarily by functional convenience rather than theoretical argument or empirical evidence, and while their later proof-of-concept analysis tunes the mapping and keeps it largely the same, the paper does not provide a general framework for establishing thresholds or mappings for any given fuzzy quantifier.

The section that follows combines Zadeh's general approach with Fernadi's guidance about one possible mapping in order to demonstrate the utility of this general method for evaluating uncertainty in a social science context.

3.2.1 Example: NLD Election results

To illustrate the value of this approach for social science research, and to simplify how it might be applied, I offer the following example evaluating the same corpus of newspaper articles from *The Irrawaddy*. Suppose in this case that a researcher sought to evaluate whether the main democratic party, the National League for Democracy, would sweep

into political power using the following proposition:

$$p \triangleq \text{“The NLD (will win/wins/won) most (of the) seats in parliament”}$$

“Most” serves as a fuzzy quantifier on the fuzzy set “seats.” In this simplified example, “NLD” is not a fuzzy set, as candidates are either members of the party or not, so the only condition to evaluate in the EDF framework is “most.” MOST (μ_M) is a condition that can be derived from external sources (i.e., what does an English speaker generically interpret “most” to mean). Following guidance specifically defining “most” (in relation to “few”) offered in Farnadi et al. (2016) yields the following:

$$\mu_M = \begin{cases} 0 & \text{if } p < 0.25; \\ \frac{(p-0.25)}{0.5} & \text{if } 0.25 \leqslant p < 0.75; \\ 1 & \text{if } p \geqslant 0.75. \end{cases}$$

In this framework, the MOST condition is derived and specified prior to the investigation of a specific proposition of interest. The external specification of the fuzzy quantifier condition has the advantage of not being biased by a particular researcher's interest or interpretation of a textual source, and being consistent across propositions and applications. At the same time, this external condition does not allow for contextual specificity or sensitivity to the unique ways in which some authors might express uncertainty. In the current example, to evaluate the proposition above textually to detect uncertainty about its possibility over time, the application of this MOST condition would come at the end of several steps:

1. Text preparation: as described above, minimal preprocessing is applied and alternate forms of “NLD” and “win” are replaced with consistent terms irrespective of tense

2. Temporal division: split the corpus into blocks by date to assess changes directly leading up to the 8 November 2015 national elections (2013, 2014, 1 Jan.–31 Oct. 2015, Nov 1–7 Nov. 2015, 8 Nov.–31 Dec. 2015)

3. Term co-occurrence:

   (a) Split text into 5-word windows

   (b) Count the frequency of co-occurrence between NLD and win within each window

   (c) Divide by the number of articles in the time block to give a per-document rate

4. Aggregate: Evaluate term co-occurrence in each time period

5. Rescale: Transform values of co-occurrence into $[0, 1]$ using $\frac{x}{x+0.05}$

Completing the above steps yields the results in Figure 22, which already indicate the expected pattern that "NLD" and "win" gain increasing association over time, peaking at the election itself. Applying the external MOST condition to these results would only serve to further scale the co-occurrence results. For the 2012 articles, for example, applying the MOST constraint above scales the possibility of the proposition to 35.4%, whereas for the lattermost time periods, 1–7 November and 8 November – 31 December 2015, it increases the possibility of the proposition to 100%. Again, these values comport with expectations (that it was completely possible for the NLD to win a majority of seats in November 2015, which they indeed did). Theoretically, however, this rescaling is unsatisfying for evaluating differences in the *uncertainty* that the articles reflect about the proposition, given that the basis for the assessment is only term co-occurrence—a particularly fickle measure where less frequent terms are concerned.

Particularly if the size of the corpus allows, generating the fuzzy quantifier constraint endogenously seems feasible and preferable. The rarity of individual fuzzy quantifiers and the under-theorization of their exchangeability, however, impose significant

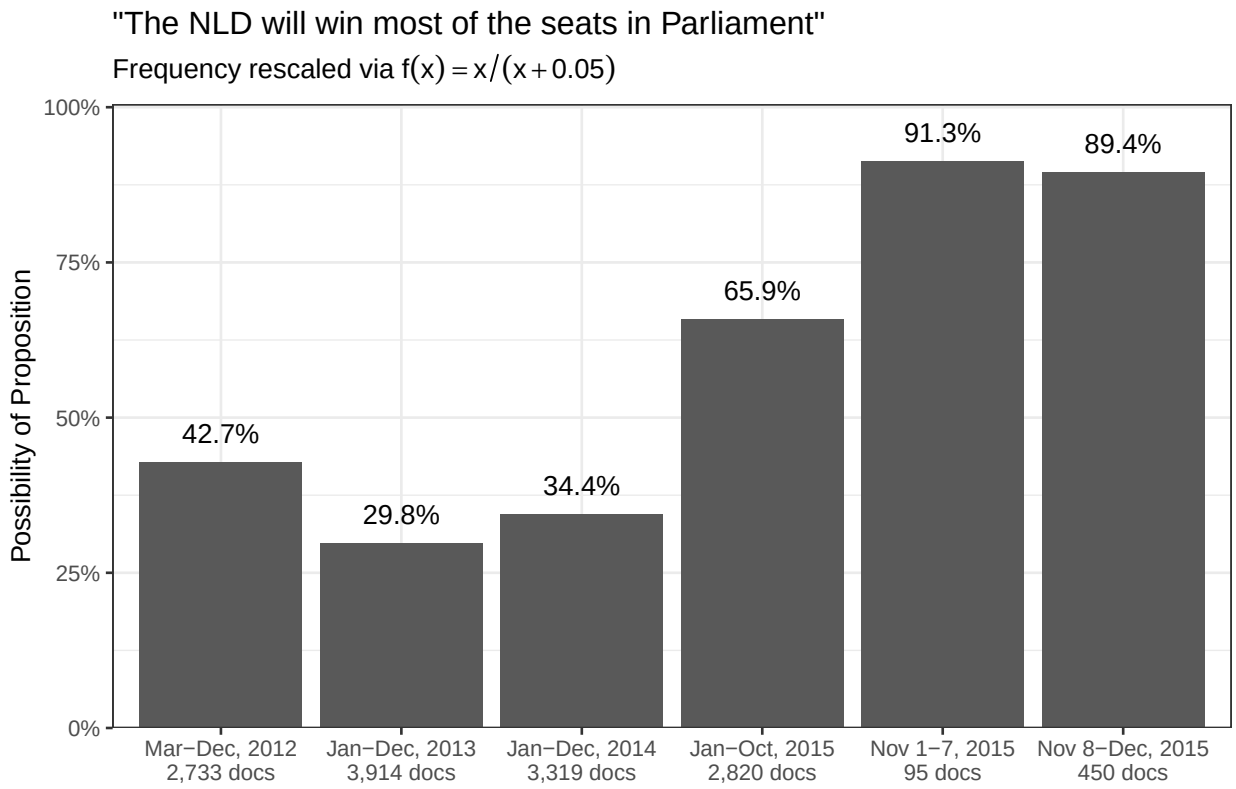

Figure 21: Association of "NLD" and "win" over time in politics articles

constraints. Figure 21 illustrates that "NLD" and "most" covary in the same way that "NLD" and "win" do within this corpus, which aligns with the general objective of specifying the fuzzy quantity of seats that the NLD was expected to "win," but the vary different scales of the associations on each axis (i.e., that the x axis is significantly smaller) underscores the need for greater distinction among fuzzy quantifiers used to detect levels of uncertainty to compensate for their rarity.

While this example serves primarily to illustrate the application of a single fuzzy quantifier constraint, this approach extends to more complex examples and propositions featuring multiple fuzzy sets. In this case, membership in the NLD is discrete, but one can easily imagine analogous propositions evaluating "liberals" or "conservatives" according to ideology, where membership is not clearly defined and party affiliation only serves as a proxy. Likewise, in this case, the "most" condition may easily map to common conceptions or metrics of political interest (i.e., a majority or a plurality), but for other propositions, such as "Most democratic countries are peaceful," simply clearing the majority threshold may not comport with theoretical expectations or useful social scientific conclusions.

3.2.2 Discussion

This method of evaluates uncertainty via "possibility" relative to the exclusively dictionary-based approaches articulated earlier. Both approaches are subject to some degree of measurement error, particularly as it pertains to evaluating which instances of fuzzy quantifier usage are relevant to the proposition ("Most democracies do not go to war" or "Between five and ten hostages were taken") versus those that are less so ("Most of the literature on democracy to date has emphasized..." or "Conflict rages between the two states"). Yet unlike dictionary-based methods that would ultimately rely on frequency of fuzzy quantifiers to determine "uncertainty," this approach has much greater flexibility and theoretical coherence. Distinctions among documents utilizing a complete

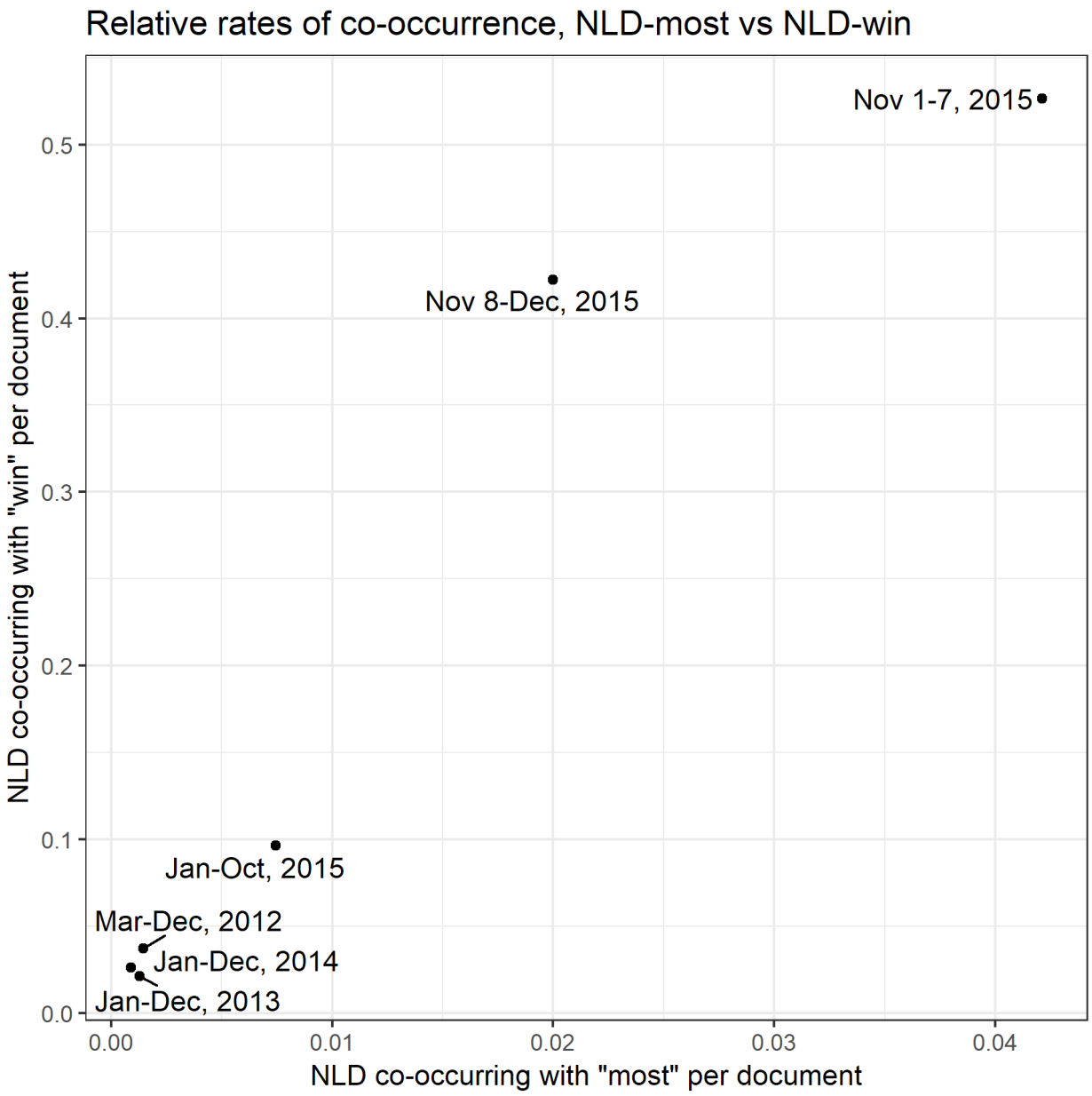

Figure 22: Association of "NLD" and "win" vs. "NLD" and "most"

fuzzy quantifier framework are more likely related to intentional argumentative structure rather than idiosyncratic and highly susceptible to the choice of aggregation rule, as in the dictionary-based cases. That said, specifying the fuzzy quantifier constraint for this approach presents a greater theoretical challenge.

The fungibility of the "most" constraint is also potentially problematic. Placing the burden on individual researchers to specify the bounds for each fuzzy quantifier they seek to evaluate in propositions seems cumbersome and raises difficulties for generalizability across research. On another hand, as Zadeh elsewhere notes, "information is a generalized constraint on the values which a variable is allowed to take" (Zadeh 2005, 2–3). In this sense, the flexibility of these constraints makes them more easily updateable with new information and new data, or to new conceptions of what might constitute set membership. Even with a requirement to generate fuzzy quantifier constraints in each new project, the fuzzy quantifier framework at least facilitates transparent and testable statements of researchers' beliefs and arguments.

A clear direction for future improvement is in finding alternative ways to specify the fuzzy quantifier constraints. As mentioned previously, one mechanism would be through a large-scale human coding exercise or survey, in which respondents evaluated possibility bounds for given fuzzy quantifier terms. Alternatively, simulations could generate thresholds with respect to particular empirical cases (these thresholds may differ across corpora, for example), to provide general guidelines for the "best" bounds, whether those are the most inclusive or the most restrictive. Furthermore, with several large-scale corpora, it would be possible to create bounds or priors for fuzzy quantifier values, much in the same way large repositories of text make it feasible to pre-fit word embeddings. Incorporating some degree of dictionary-based detection of fuzzy quantifier terms in relevant articles might allow for endogenously generating the fuzzy quantifier constraint from within a particular corpus or corpora. For example, within

these articles from *The Irrawaddy*, usage of "most" relative to other types of proportion-specifying fuzzy quantifiers could serve as a basis to endogenously generate a "most" condition, rather than having a researcher-generated condition imposed.

4 Conclusions & Extensions

This paper provides a preliminary sense of the ways in which fuzzy quantifiers and fuzzy logic can undergird attempts to estimate uncertainty from text for social science applications. Aside from the pre-labeled corpora used previously for comparison to subjectivity analyses, the dictionary-based and broader fuzzy logic framework articulated here can apply to a multitude of differing data formats relevant for social science questions.

4.1 Additional Applications

The applications discussed above presume textual data coming from published research, whether in the form of papers or monographs. The methods, however, apply just as easily to other types of textual or qualitative data in social science contexts. For interview or archival data, fuzzy quantification with respect to propositions in particular could provide more accurate assessments of beliefs and uncertainty, particularly for respondents or authors who are not trained in social science methods and frameworks or statistics, or for sources from disparate time periods or locations where validation and confirmation is challenging or impossible.

Likewise, these methods could apply to open-ended survey responses, where evaluating subjects' uncertainty about questions or claims has traditionally presented a challenge. For example, survey responses with both open- and closed-ended questions could conduct the fuzzy quantifier procedure described above on textual data and correlate with levels of uncertainty expressed elsewhere for validity and scaling. A preliminary assessment using the 2016 American National Election Studies (ANES) time series data

indicates that fewer than 5% of responses (per question on average) to closed-ended questions are "I don't know," but that these responses are weakly positively associated with fuzzy quantifier usage in corresponding open-ended questions (American National Election Studies 2016). More long-form open-ended survey questions and other cues for uncertainty could further confirm and refine the relationship between fuzzy quantifier terminology and construct usage and uncertainty expression.

With published works, conversely, an additional dimension to consider is editorial effects on uncertain expression. A preliminary investigation of Wikipedia articles indicated no significant relationship between fuzzy quantifier usage and the number of editors or edits for a given page (see Appendix), but for academic journal publications or other venues, editorial style or editorial board fixed effects may dictate norms around either fuzzy quantifier term usage or standards and bounds that define particular fuzzy quantifiers. Further investigating uncertainty levels not only at the document and corpus level, but also according to author and editorial staff, would further illuminate the locus from which uncertainty expression conventions arise.

4.2 Corpus Selection

Fuzzy logic and fuzzy set frameworks have numerous possible applications in other aspects of text analysis that are particularly relevant to the social sciences. It may, in fact, be useful in structuring the way text analysis is performed. While in other disciplines, corpora may be more readily defined by the task (e.g., medical journal articles evaluating breast cancer treatment prognoses, or the collected works of Ayn Rand), establishing the scope of corpora for many social and political science questions has traditionally been done by argumentation. Fuzzy set logic could apply to the inclusion of documents in a corpus, providing not only a set of possible corpus members but also establishing degrees of membership.

4.3 Multi-Lingual Applications

Evaluating uncertainty across multiple languages presents its own unique challenges (Bentz and Alikaniotis 2016). Languages differ in their conceptions of measurement and time, which can complicate efforts to precisely or quantitatively characterize arguments within the text. Languages like Amharic are in fact characterized by their distinct ambiguity, particularly in spoken language, but also in terms of word choice and order (Amare 2001), whereas languages like Burmese would generate problems for leveraging precisely the theoretical distinctions this paper attempts to articulate. In Burmese, by way of example, no distinction exists between probability and possibility; likewise, saying an event or condition is "likely" is analogous to saying it is "possible." These distinctions are further complicated, depending on the context or type of text, by culturally dictated mechanisms for maintaining safety via vagueness or omission (e.g., if asked a sensitive question about a political situation, one might answer in Burmese, "could be" or "it's possible"). Certain kinds of fuzzy quantification, however, are more consistent across language frames despite, or because of, their indeterminate reference sets. "Many," "most," or "few," are less culturally moored assessments of uncertainty than typical subjectivity terms, and therefore may serve as a more consistent basis for establishing cross-language classification and measurement schema for uncertainty.

References

Akkaya, Cem, Janyce Wiebe, and Rada Mihalcea. 2009. "Subjectivity Word Sense Disambiguiation." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:* 190–199.

Amare, Getahun. 2001. "Towards the Analysis of Ambiguity in Amharic." *Journal of Ethiopian Studies* 34 (2): 35–56.

American National Election Studies. 2016. "THE ANES GUIDE TO PUBLIC OPINION AND ELECTORAL BEHAVIOR." *University of Michigan, Center for Political Studies [producer and distributor].* Ann Arbor, MI. `www.electionstudies.org`.

Atanassov, Krassimir, and George Gargov. 1998. "Elements of Intuitionistic Fuzzy Logic, Part I." *Fuzzy Sets and Systems* 95:39–52.

Auger, Alain, and Jean Roy. 2008. "Expression of Uncertainty in Linguistic Data." *Proceedings of the 11th International Conference on Information Fusion (FUSION):* 1860–1867.

Barwise, Jon, and Robin Cooper. 1981. "Generalized Quantifiers and Natural Language." *Linguistics and Philosophy* 4:159–219.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53 (2): 495–513.

Bentz, Christian, and Dimitrios Alikaniotis. 2016. "The Word Entropy of Natural Languages." *arXiv.* `arxiv.org/1606.06996`.

Breck, Eric, and Claire Cardie. 2014. "Opinion Mining and Sentiment Analysis." In *The Oxford Handbook of Computational Linguistics,* 2nd ed., edited by Ruslan Mitov. Oxford: Oxford University Press.

Cabrerizo, F., R. Al-Hmouz, A. Morfeq, A. Balamash, M. Martìnez, and E. Herrera-Viedma. 2017. "Soft Consensus Measures in Group Decision Making Using Unbalanced Fuzzy Linguistic Information." *Soft Computing* 21:3037–3050.

Chen, Chaomei, Ming Song, and Go Eun Heo. 2017. "A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty." *arXiv.* `arxiv.org/1710.08327`.

Conde-Clemente, Patricia, Jose M. Alonso, Èldman O. Nunes, Angel Sanchez, and Gracian Trivino. 2017. "New Types of Computational Perceptions: Linguistic Descriptions in Deforestation Analysis." *Expert Systems with Applications* 85:46–60.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26 (2): 168–189.

Dhar, Mamoni. 2013. "Cardinality of Fuzzy Sets: An Overview." *International Journal of Energy, Information and Communications* 4 (1): 15–22.

Dìaz-Hermida, F., A. Bugarìn, and S. Barro. 2003. "Definition and Classification of Semi-Fuzzy Quantifiers for the Evaluation of Fuzzy Quantified Sentences." *International Journal of Approximate Reasoning* 34:49–88.

Dragos, Valentina. 2013. "An Ontological Analysis of Uncertainty in Soft Data." *16th International Conference on Information Fusion (FUSION).* `https://www.researchgate.net/publication/261316990_An_ontological_analysis_of_uncertainty_in_soft_data`.

Druzdzel, Marek. 1989. "Verbal Uncertainty Expressions: Literature Review." *Technical Report: CMU-EPP 1990-03-02.*

Farkas, Richàrd, Veronika Vincze, György Mòra, Jànos Csirik, and György Szarvas. 2010. "The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text." *Conference on Computational Natural Language Learning (CoNLL).* `http://www.aclweb.org/anthology/W10-3001`.

Farnadi, Golnoosh, Stephen H. Bach, Marjon Blondeel, Marie-Francine Moens, Lise Getoor, and Martine De Cock. 2016. "Statistical Relational Learning with Soft Quantifiers." In *Inductive Logic Programming 25th International Conference Selected Papers,* edited by Katsumi Inoue, Hayato Ohwada, and Akihiro Yamamoto, 61–75. Springer International.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Jean, Pierre-Antoine, Sèbastien Harispe, Sylvie Ranwez, Patrice Bellot, and Jacky Montmain. 2016. "Uncertainty Detection in Natural Language: A Probabilistic Model." *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics.*

Khoury, Richard, Fakhreddine Karray, and Mohamed Kamel. 2008. "Domain Representation Using Possibility Theory: An Exploratory Study." *IEEE Transactions on Fuzzy Systems* 16 (6): 1531–1541.

Lassiter, Daniel, and Noah B. Goodman. 2017. "Adjectival Vagueness in a Bayesian Model of Interpretation." *Synthese* 194:3801–3836.

Li, Xiujun, Wei Gao, and Jude W. Shavlik. 2014. "Detecting Semantic Uncertainty by Learning Hedge Cues in Sentences Using an HMM." *Proceedings of Workshop on Semantic Matching in Information Retrieval:* 30–37.

Lindley, Dennis V. 1987. "The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems." *Statistical Science* 2 (1): 17–24.

Liu, Yaxin, and Etienne E. Kerre. 1998. "An Overview of Fuzzy Quantifiers I: Interpretations (Invited Review)." *Fuzzy Sets and Systems* 95:1–21.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Miller, Ben, Ayush Shrestha, Jason Derby, Jennifer Olive, Karthikeyan Umapathy, Fuxin Li, and Yanjun Zhao. 2013. "Digging into Human Rights Violations: Data Modelling and Collective Memory." *2013 IEEE International Conference on Big Data:* 37–45.

Ramos-Soto, A., and M. Pereira-Farina. 2017. "On Modeling Vagueness and Uncertainty in Data-to-Text Systems through Fuzzy Sets." *arXiv.* `arxiv.org/1710.10093`.

Roos, Nico. 1990. "How to Reason with Uncertain Knowledge." *Proceedings of the 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems:* 403–412.

Rubin, Victoria L., Elizabeth D. Liddy, and Noriko Kando. 2006. "Certainty Identification in Texts: Categorization Model and Manual Tagging Results." Chap. 7 in *Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series,* edited by James Shanahan, Yan Qu, and Janyce Wiebe, 20:61–76. Springer.

Saif, Hassan, Miriam Fernandez, Yulan He, and Harith Alani. 2014. "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter." *Language Resources Evaluation Conference (LREC) Proceedings.* `http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf`.

"Szeged Uncertainty Corpus." 2010. `http://rgai.inf.u-szeged.hu/index.php?lang=en&page=uncertainty`.

Szmidt, Eulalia, and Janusz Kacprzyk. 2001. "Entropy for Intuitionistic Fuzzy Sets." *Fuzzy Sets and Systems* 118:467–477.

Thomson, Judi, Beth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. 2005. "A Typology for Visualizing Uncertainty." *Conference on Visualization and Data Analysis, IS&T/SPIE Symposium on Electronic Imaging.*

Vincze, Veronika. 2014. "Uncertainty Detection in Natural Language Texts." *Dissertation, Research Group on Artificial Intelligence and University of Szeged.* `http://doktori.bibl.u-szeged.hu/2291/1/Vincze_Veronika_tezis.pdf`.

Wygralak, Maciej. 1998. "Vagueness and its Representations: A Unifying Look." *Mathware and Soft Computing* 5:121–131.

Zadeh, Lotfi A. 1983. "A Computational Approach to Fuzzy Quantifiers in Natural Languages." *Computers and Mathematics with Applications* 9 (1): 141–184.

———. 2005. "Toward a Generalized Theory of Uncertainty (GTU)–An Outline." *Information Sciences* 172:1–40.

———. 2008. "Is there a need for fuzzy logic?" *Information Sciences* 178:2751–2779.

Zhai, Daoyuan, and Jerry M. Mendel. 2011. "Uncertainty Measures for General Type-2 Fuzzy Sets." *Information Sciences* 181:503–518.

A Word Lists: Fuzzy Quantifiers, Precise Quantifiers, Numbers

B Fuzzy Quantifiers and Edits on Wikipedia

One possible concern with the use of Wikipedia-trained word embeddings is an uneven distribution of fuzzy quantifier usage around particular topics, or arising from particular editorial styles and systems, that mean associations derived from the corpus are biased, particularly for subject matter of interest to social scientists. For example, if all articles on contentious topics have a larger number of editors or a larger number of edits, the usage of fuzzy quantifiers may proliferate (or alternatively, bottom out), and may be measuring polarization around the topic rather than uncertainty, strictly speaking, or may simply encompass imprecision of language due to a large number of contributors, which is likely theoretically different from substantive uncertainty.

The figures below provide a simple, superficial illustration for whether these types of relationships are likely to hold. Each figure uses text from articles across three broad subjects—quantum mechanics, the "War on Terror," and World War II—to assess fuzzy quantifier usage, with each serving as a different example of a "contentious" subject matter. Quantum mechanics provides an example of a topic that is both "scientific" and features some contention (with classical mechanics) for a (likely) small population of editors; the "War on Terror," in contrast, is a more contemporary topic with potentially greater polarity among editors; and World War II is one of the largest topics (by number of articles and stubs), featuring both a series of short, timeline-like stubs and longer social-historical analyses that span the spectrum of contention and uncertainty. Figure 23 indicates that, much like the other text sources examined here, fuzzy quantifiers are fairly sparse across a variety of topics, ranging from "objective" scientific subject matter to explicitly contentious issues such as the War on Terror. While not definitive, this cursory investigation indicates that fuzzy quantifier usage on Wikipedia may not differ

substantially from what we should "expect" in other published or edited media formats.

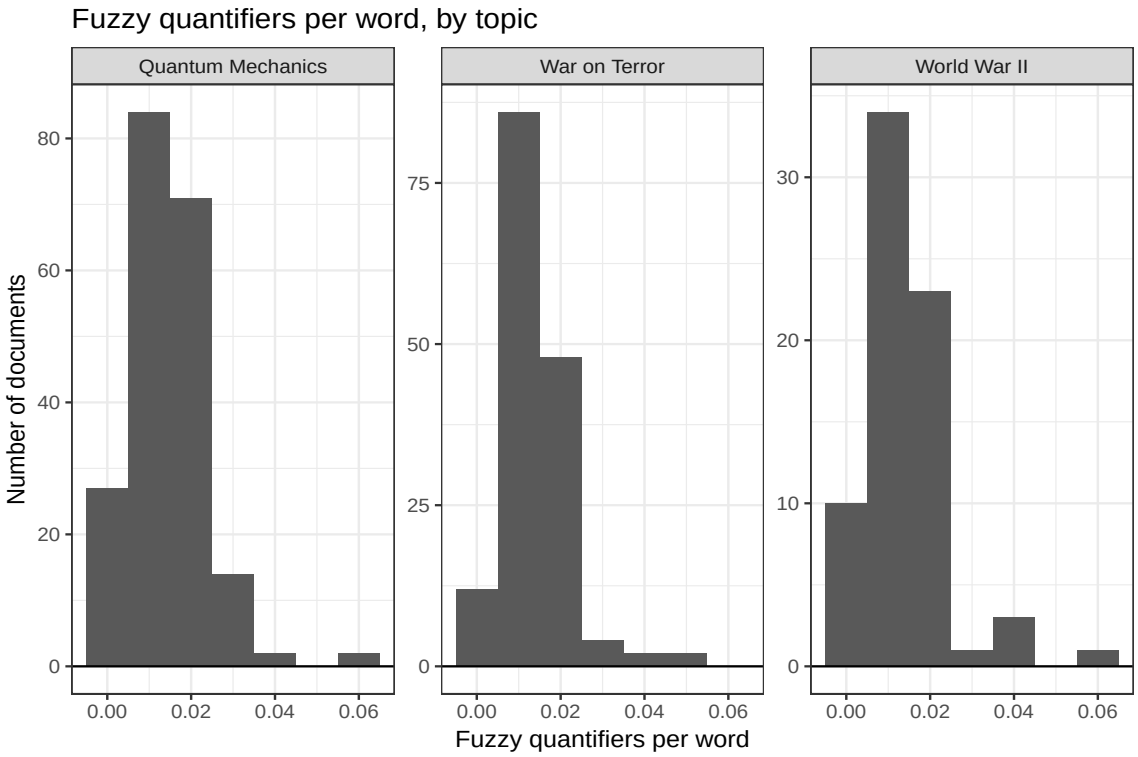

Figure 23: Wikipedia: Fuzzy Quantifier Usage by Word across 3 Example Topics

Likewise, an examination of fuzzy quantifier usage by word relative to the number of editors and the number of edits that articles and stubs within a given topic receive indicates very a very slight, but not systematic, positive association with fuzzy quantifier usage. That these broad relationships appear largely the same (largely characterized by noise), irrespective of the "objectivity" of the subject matter in these examples, is encouraging for the prospect that fuzzy quantifier usage on Wikipedia in general should provide an adequate baseline.

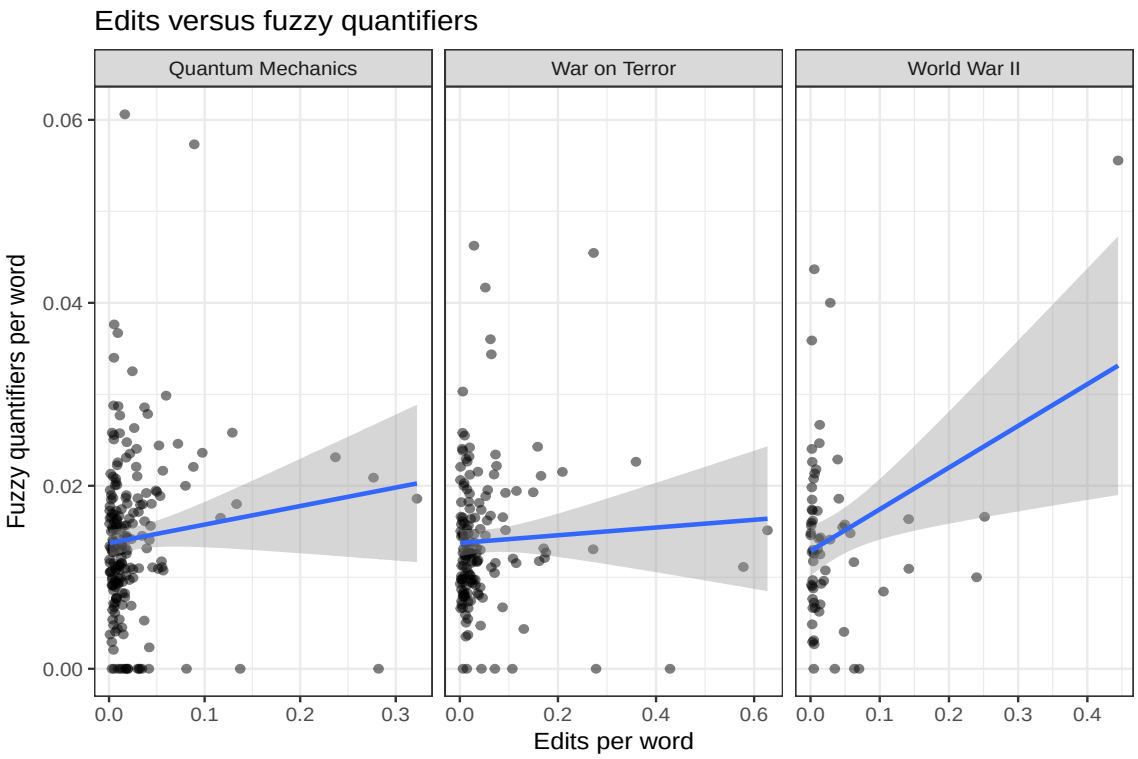

Figure 24: Wikipedia: Fuzzy Quantifier Usage vs. Number of Edits by Topic, Examples

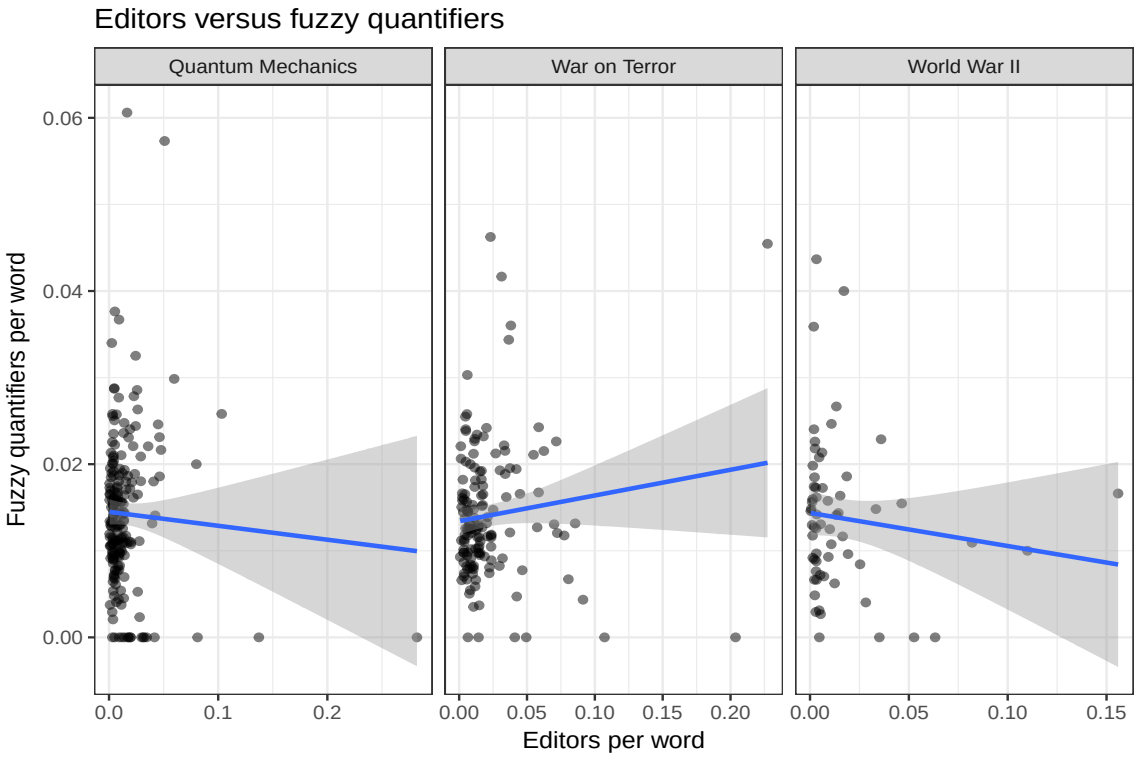

Figure 25: Wikipedia: Fuzzy Quantifier Usage by Number of Editors by Topic, Examples

Quantifier	Source
about	Liu and Kerre (1998)
additional	Conde-Clemente et al. (2017)
all	Zadeh (1983)
almost	Zadeh (1983)
another	BioScope (hand-coded)
any	Conde-Clemente et al. (2017)
approximate	Zadeh (1983)
around	BioScope (hand-coded)
as many	Zadeh (1983)
at least	Zadeh (1983)
between	BioScope (hand-coded)
broad	BioScope (hand-coded)
certain	BioScope (hand-coded)
couple	BioScope (hand-coded)
few	Zadeh (1983), Farnadi et al. (2016)
frequently	BioScope (hand-coded)
great	BioScope (hand-coded)
half	Zadeh (1983)
high	Conde-Clemente et al. (2017)
increased	BioScope (hand-coded)
less	Zadeh (1983)
low	Conde-Clemente et al. (2017)
many	Zadeh (1983)
minority	Liu and Kerre (1998)
more	Zadeh (1983)
most	Zadeh (1983), Farnadi et al. (2016)
multiple	BioScope (hand-coded)
narrow	BioScope (hand-coded)
nearly	BioScope (hand-coded)
no	Liu and Kerre (1998)
none	Zadeh (1983)
often	Zadeh (1983)
rarely	BioScope (hand-coded)
reduced	BioScope (hand-coded)
seldom	BioScope (hand-coded)
several	Zadeh (1983)
slight	BioScope (hand-coded)
some	Zadeh (1983)
twice	BioScope (hand-coded)
ubiquitous	BioScope (hand-coded)
within	BioScope (hand-coded)

Table 1: Fuzzy Quantifiers

Word	Source
ampere	https://en.wikipedia.org/wiki/International_System_of_Units
kelvin	https://en.wikipedia.org/wiki/International_System_of_Units
mole	https://en.wikipedia.org/wiki/International_System_of_Units
candela	https://en.wikipedia.org/wiki/International_System_of_Units
radian	https://en.wikipedia.org/wiki/International_System_of_Units
steradian	https://en.wikipedia.org/wiki/International_System_of_Units
newton	https://en.wikipedia.org/wiki/International_System_of_Units
pascal	https://en.wikipedia.org/wiki/International_System_of_Units
joule	https://en.wikipedia.org/wiki/International_System_of_Units
coulomb	https://en.wikipedia.org/wiki/International_System_of_Units
farad	https://en.wikipedia.org/wiki/International_System_of_Units
ohm	https://en.wikipedia.org/wiki/International_System_of_Units
siemens	https://en.wikipedia.org/wiki/International_System_of_Units
weber	https://en.wikipedia.org/wiki/International_System_of_Units
tesla	https://en.wikipedia.org/wiki/International_System_of_Units
henry	https://en.wikipedia.org/wiki/International_System_of_Units
lumen	https://en.wikipedia.org/wiki/International_System_of_Units
lux	https://en.wikipedia.org/wiki/International_System_of_Units
becquerel	https://en.wikipedia.org/wiki/International_System_of_Units
gray	https://en.wikipedia.org/wiki/International_System_of_Units
sievert	https://en.wikipedia.org/wiki/International_System_of_Units
katal	https://en.wikipedia.org/wiki/International_System_of_Units
short ton	https://www.britannica.com/science/British-Imperial-System
long ton	https://www.britannica.com/science/British-Imperial-System
hundredweight	https://www.britannica.com/science/British-Imperial-System
short hundredweight	https://www.britannica.com/science/British-Imperial-System
long hundredweight	https://www.britannica.com/science/British-Imperial-System
grain	https://www.britannica.com/science/British-Imperial-System
pennyweight	https://www.britannica.com/science/British-Imperial-System
gill	https://www.britannica.com/science/British-Imperial-System
fluid dram	https://www.britannica.com/science/British-Imperial-System
minim	https://www.britannica.com/science/British-Imperial-System
peck	https://www.britannica.com/science/British-Imperial-System
rod	https://www.britannica.com/science/British-Imperial-System
square rod	https://www.britannica.com/science/British-Imperial-System
acre-foot	https://www.britannica.com/science/British-Imperial-System
board foot	https://www.britannica.com/science/British-Imperial-System
cord	https://www.britannica.com/science/British-Imperial-System

Table 2: Units of measurement: Obscure

Word	Source
meter	https://en.wikipedia.org/wiki/International_System_of_Units
metre	different version of ”meter”
kilogram	https://en.wikipedia.org/wiki/International_System_of_Units
gram	root of ”kilogram”
second	https://en.wikipedia.org/wiki/International_System_of_Units
minute	extension of ”second”
hour	extension of ”second”
day	extension of ”second”
week	extension of ”second”
year	extension of ”second”
decade	extension of ”second”
century	extension of ”second”
millenium	extension of ”second”
hertz	https://en.wikipedia.org/wiki/International_System_of_Units
watt	https://en.wikipedia.org/wiki/International_System_of_Units
volt	https://en.wikipedia.org/wiki/International_System_of_Units
degree Celsius	https://en.wikipedia.org/wiki/International_System_of_Units
liter	https://en.wikipedia.org/wiki/International_System_of_Units#Non-SI_units_accepted_for_use_with_SI
litre	https://en.wikipedia.org/wiki/International_System_of_Units#Non-SI_units_accepted_for_use_with_SI
tonne	https://en.wikipedia.org/wiki/International_System_of_Units#Non-SI_units_accepted_for_use_with_SI
hectare	https://en.wikipedia.org/wiki/International_System_of_Units#Non-SI_units_accepted_for_use_with_SI
ton	https://www.britannica.com/science/British-Imperial-System
pound	https://www.britannica.com/science/British-Imperial-System
ounce	https://www.britannica.com/science/British-Imperial-System
dram	https://www.britannica.com/science/British-Imperial-System
stone	https://www.britannica.com/science/British-Imperial-System
gallon	https://www.britannica.com/science/British-Imperial-System
quart	https://www.britannica.com/science/British-Imperial-System
pint	https://www.britannica.com/science/British-Imperial-System
fluid ounce	https://www.britannica.com/science/British-Imperial-System
bushel	https://www.britannica.com/science/British-Imperial-System
nautical mile	https://www.britannica.com/science/British-Imperial-System
mile	https://www.britannica.com/science/British-Imperial-System
furlong	https://www.britannica.com/science/British-Imperial-System
fathom	https://www.britannica.com/science/British-Imperial-System
yard	https://www.britannica.com/science/British-Imperial-System
foot	https://www.britannica.com/science/British-Imperial-System
inch	https://www.britannica.com/science/British-Imperial-System
square mile	https://www.britannica.com/science/British-Imperial-System
acre	https://www.britannica.com/science/British-Imperial-System
square yard	https://www.britannica.com/science/British-Imperial-System
square foot	https://www.britannica.com/science/British-Imperial-System
square inch	https://www.britannica.com/science/British-Imperial-System
cubic yard	https://www.britannica.com/science/British-Imperial-System
cubic foot	https://www.britannica.com/science/British-Imperial-System
cubic inch	https://www.britannica.com/science/British-Imperial-System

Table 3: Units of Measurement: Common

Word	Type
zero	integer
one	integer
two	integer
three	integer
four	integer
five	integer
six	integer
seven	integer
eight	integer
nine	integer
ten	integer
eleven	integer
twelve	integer
thirteen	integer
fourteen	integer
fifteen	integer
sixteen	integer
seventeen	integer
eighteen	integer
nineteen	integer
twenty	tens
thirty	tens
forty	tens
fifty	tens
sixty	tens
seventy	tens
eighty	tens
ninety	tens
hundred	above tens
thousand	above tens
million	above tens
billion	above tens
trillion	above tens

Table 4: Number words