Scare Quotes as Markers of "Questionable" Word Usages and Misalignment in Conversation: An Annotation Study

Aina Garí Soler[1,2] **Juan Carlos Zevallos Huaco**[3]
Matthieu Labeau[4] **Chloé Clavel**[1]
[1]INRIA, Paris [2]Miles Team, LAMSADE, Université Paris Dauphine-PSL, Paris, France
[3]Independent Researcher [4]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
aina.gari-soler@inria.fr

Abstract

Scare quotes are a subtle yet powerful device: they can mark irony, distance, or disagreement about word meaning or lexical choices. We present a large-scale manual annotation of quoted word usages focused on the scare versus non-scare quote distinction as well as on their role in managing (mis)alignment in conversation. Our analysis reveals that scare quotes can mark problematic word usages, and they are often used to contest or criticize other speakers' word choices. However, non-scare, meta-linguistic usages of quotes are also often involved in explicit efforts toward lexico-semantic alignment.

Keywords: Scare quotes, Quotation marks, Corpus annotation, Dialog, Conversational alignment, Lexical semantics

1. Introduction

Quotation marks (QMs, " ", ' ') are a typographic device with a wide range of uses in written language. In English, they often serve to directly quote someone else's words (*He said "I'm hungry"*) or mark meta-linguistic mentions of a word (*The word "dialogue" is of Greek origin*). Among their multiple functions, QMs are also added to indicate that a word or a phrase should not be taken entirely at face value (*The "expert" couldn't answer a single question*). Such usages are typically referred to as "scare quotes" (Gutzmann and Stei, 2011). Scare quotes serve as a cue to the reader to direct their attention to a word that is used in a non-standard way. The nature of this "non-standardness"[1] can be very varied: they can signal irony, skepticism, the lack of a better term, or an unusual or questionable expression.

In this paper, we approach scare quotes as explicit markers of potentially problematic word usages (Garí Soler et al., 2025a) in conversation; that is, word usages that may cause a misunderstanding or disagreement of a semantic nature. From this perspective, scare quotes are particularly interesting: they show the writer's acknowledgment that a word may be problematic and their anticipation of a possible misalignment in meaning between interlocutors, as well as an attempt to mitigate it by warning the reader that a word may require careful interpretation. By pointing the reader directly to the potential problem source, interlocutors can trigger meta-linguistic and repair-oriented discussions aimed at preventing misunderstandings.

While there is an extensive body of linguistic research on different types of quoting, most works are theoretical in nature (see Section 2). Empirical work remains limited and has mainly been focused on quotation as a means for citing someone. Despite the widespread use of scare quotes, existing discussions based on real data rely on small-scale, non-public datasets and do not explore the interactional and alignment-oriented role of scare quotes (Schneider, 2002; Nacey, 2012; Nádraská, 2022; Xiong and Robles, 2023), focusing instead on other specific aspects such as stigmatization (Walker et al., 2025).

We fill this gap by introducing the first large-scale, publicly available annotation of functions of quotation.[2] We collect and annotate QM usages from online written dialog, specifically from social media debates, which allows us to investigate the function of scare quoting in interaction. Our contributions are as follows:

- We provide the first annotated corpus of quoted word and phrase usages focused on the scare versus non-scare quotes distinction: **SQuiC** (Scare Quotes in Conversation);
- we develop an annotation schema of quote functions grounded both in the existing literature and in real data, designed to be general-purpose and easily adaptable to other corpora, which is facilitated by the release of our annotation guidelines;
- we analyze the characteristics of quoted items, highlighting their role in causing, indicating and

[1]Incidentally, these scare quotes serve to acknowledge that this term is not in the dictionary.

[2]Annotations, guidelines and code can be found at https://github.com/ainagari/scare-quotes

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11834–11851
11–16 May 2026.

repairing lexico-semantic misalignment in conversation (Garí Soler et al., 2023);

- we evaluate the ability of large language models (LLMs) to annotate quoted usages with the scare versus non-scare distinction.

The paper is structured as follows. Section 2 reviews theoretical and empirical work on QMs and scare quoting. In Section 3 we introduce our data collection and annotation process and our proposed classification schema. Section 4 presents annotation statistics with attention to lexico-semantic alignment, analyzing the lexical properties of quoted items; and reports inter-annotator agreement and LLM classification results. Finally, Section 5 concludes with a discussion of our findings and directions for future work.

2. Background

2.1. Quotation Marks and Quoting

The majority of research on quotation has been theoretical in nature, often centered on the debate of whether its function is primarily semantic or pragmatic (Recanati, 2001; Predelli, 2003; De Brabanter, 2010; Gutzmann and Stei, 2011; De Brabanter, 2017). A number of classifications of QM functions have been proposed, varying in scope and granularity (Cappelen and Lepore, 1997, 2003). For example, Gutzmann and Stei (2011) distinguish between five different types: *pure quotation* (a metalinguistic use of a word), *direct quotation* (reproducing someone else's words), *mixed quotation* (a partial reproduction of someone else's words embedded in the author's own wording, as in *He said that he was "hungry"*), *scare quotes* (marking non-literal, ironic, or distancing usages) and *emphatic quotes* (giving emphasis to a word). These categories are not fully mutually exclusive: mixed quotation may also convey distancing or irony, overlapping with scare quoting (De Brabanter, 2010). Recanati (2001), instead, distinguishes two main classes: *closed* and *open quotation*. The former includes pure and direct quotation, whereas open quotation would include all other types.

These taxonomies are grounded on real examples, but are not established based on large-scale corpus studies. Despite the heterogeneous set of functions of QMs, there is relatively little empirical work examining how they are used in practice. Most corpus-based research has focused on direct quotation (Musolff, 2015; Ricci and Rossari, 2018), often in journalistic or social media contexts (Haapanen and Perrin, 2017; Kula and Grzelka, 2022). Similarly, in Natural Language Processing, works on quotation detection have concentrated on identifying cited text (Scheible et al., 2016; Papay and Padó, 2019).

While QMs can have oral correlates expressed with intonation (Schlechtweg and Härtl, 2020) or air quotes (Cirillo, 2019), in this paper we focus our attention on written text.

2.2. Scare Quoting

Scare quotes (hereafter SQ) are a particular use of QMs representing "a warning to the reader that there is something unusual or dubious (in the opinion of the writer) about the quoted word or phrase" (McArthur et al., 1992).[3] They can reflect skepticism, irony, distance, novelty, tentativeness or a non-standard or inappropriate usage, and can be seen as metadiscoursal devices that signal some interaction between writer and readers, or between a writer and their text (Nádraská, 2022; Hyland, 2010). They can often replace expressions such as "so-called", "roughly speaking" or "so to speak" (Predelli, 2003). SQ (or similar, partially-overlapping concepts) have also been referred to as "nonstandard quotes" (Schneider, 2002) or "perverted commas" (Dillon, 1988). They are often discouraged in prescriptive style guides such as the Chicago Manual of Style (University of Chicago Press, 2017), yet they are used by both proficient and novice writers (Schneider, 2002); they are useful for non-native speakers when they do not manage to express exactly what they want (Nacey, 2012); and it has been suggested that their usage may have increased in the last years (McCullagh, 2017). Despite their ubiquity in everyday language, early works on quotation often ignored this function of quotes, focusing instead on direct and pure quotation (Cappelen and Lepore, 1997).

SQ can have multiple functions. Based on existing handbooks, Dillon (1988) proposes a taxonomy which includes *shudder quotes* (for slang or inappropriate words) and *figurative senses*, among others. However, when applying this taxonomy to real data, Schneider (2002) found categories to overlap, and uses that were not accounted for. Other researchers emphasize the use of SQ for expressing insecurity about word choice (Nacey, 2012) or to mark concepts that are controversial (Nádraská, 2022).

Corpus research on SQ is scarce. Nacey (2012) investigates the use of SQ in essays by novice writers. The author proposes her own taxonomy revolving around "secure" versus "insecure" scare quotes. Insecure SQ include phenomena such as substitutions (e.g., using a general word instead of a more specific one; borrowing from another language, etc.). Nádraská (2022) instead explores SQ in hard news. She focuses on quotes that express some attitude towards the quoted passage, with emphasis on whether the words originate from an in-

[3] Cited from De Brabanter (2010).

ternal voice or they can be attributed to an external voice. The latter would be a combination of mixed quotation with some attitude. However, these annotations, including Schneider (2002)'s, are not publicly available and were carried out at a small scale (they each consider less than 100k words, and the three studies together would amount to about 300 SQ cases). The developed taxonomies are very specific to the phenomenon or data being studied and there is no agreement study, since the goal was to have a classification that stirred further discussion, rather than a definitive, reliable annotation (Schneider, 2002; Nacey, 2012). Xiong and Robles (2023) carry out a manual annotation of quoted usages which is done on a larger scale, but restrict the data to comments from two online videos, and the categories they propose are also very specific to their data (e.g., "political sloganizing").

The use of SQ has also sparked interest in the healthcare domain. Although QMs can be seen as a means for faithfully reproducing a patient's words in electronic health records, they are sometimes used in a way that stigmatizes the patients instead (P. Goddu et al., 2018). Some works have found that QMs are used more often with black patients than with white patients (Beach et al., 2021; Piscitello et al., 2025), although without making a formal distinction between SQ and other functions of quotes. Walker et al. (2025) carry out an annotation of quotes with multiple annotators and public guidelines, but it is restricted to stigmatization.

Our work is thus the first large-scale, empirical study of SQ with a broadly applicable classification schema and also the first to focus on a conversational setting.

3. Data Collection and Annotation

This section presents the data used (Section 3.1) and our annotation schema and procedure (Section 3.2).

3.1. Data Collection

We base our study on *Winning Arguments* (Tan et al., 2016), an English corpus compiled from the subreddit r/ChangeMyView between 2013 and 2015. In this forum, users post a controversial opinion and invite others to challenge it, which results in debate-like exchanges. One of the main reasons for selecting this dataset is that it has been enriched with **Word Meaning Negotiation (WMN)** annotations as part of the NeWMe corpus (Garí Soler et al., 2025). WMN sequences are stretches of conversation involving a problematic word usage (a **trigger**) that prompts a speaker to question said usage (with an **indicator**), often leading to an explicit meta-linguistic discussion (the **negotiation**) about word meaning. This allows us to enrich our analysis with a perspective on lexico-semantic alignment, as well as to facilitate further studies on WMN and their relationship with SQ. See Appendix A for examples of WMN sequences.

Another reason for selecting this dataset, besides its conversational nature, is that a debate setting is particularly conducive to quoting, as speakers often want to ensure that they are using terms under the same definitions, or engage in meta-linguistic discussions of word meaning. Additionally, social media interactions are interesting because we expect to encounter more SQ in such an informal setting, as opposed to more formal contexts where their use may be prescriptively discouraged.

Following Garí Soler et al. (2025), we restrict our scope to the 586 conversations with no deleted authors or posts. This consists of approximately 3 million words and 27k posts. We focus on short quoted expressions of up to three words. This restriction is motivated by our interest in problematic word usages and SQ, which typically involve individual words or short phrases, while longer quoted spans are more likely to contain citations. Limiting the length thus increases the chances of capturing instances of SQ while ensuring that the retained non-scare quote examples are comparable and potentially confusable with them. We extract candidates containing double (" ") or single (' ') QMs using regular expressions. We ignore all Reddit-formatted citations of previous messages to include only matches from original posts rather than cited material. This search yields 9,233 matches from 5,200 distinct posts (19% of all posts).

3.2. Annotation

In this section we describe the composition of the annotated sample and the annotation stages (Section 3.2.1) and present the classification schema (Section 3.2.2) and the inter-annotator agreement (Section 3.2.3).

3.2.1. Annotation sample and process

We manually annotated a total of 2,500 instances. The annotation was carried out by an author of this paper using the LabelStudio platform.[4] For each quotation mark usage (QMU), the annotator was shown the post containing the quote as well as the preceding message chain and all subsequent replies. Annotation was conducted in two stages.

In the first stage, 1,000 randomly sampled instances were annotated with a coarse-grained distinction (SQ versus non-scare quotes, NSQ) and a preliminary set of fine-grained labels derived from

[4] https://labelstud.io/.

	1-grams	2-grams	3-grams	Total
Total	1,496	669	335	2,500
Unique	1,089	605	319	2,013
Double QM	1,286	590	314	2,190
Single QM	210	79	21	310
WMN	485	149	56	690
Triggers	52	17	5	74
Indicators	160	37	19	216
Negotiation	276	96	33	405

Table 1: Composition of the annotated sample.

the literature. This stage served to enable an inductive analysis of the data for developing the final annotation schema and drafting the annotation guidelines.

In the second stage, after finalizing the annotation schema and guidelines, we first revisited the 1,000 instances to complete and, if necessary, correct the annotation. Then we selected 1,500 more QMUs, ensuring (1) full coverage of the conversational phenomena annotated in NeWMe[5] and (2) maximal diversity of threads and quoted passages. This was enforced by including at least two QMUs from each available thread, when possible; prioritizing quoted passages that were not yet in the sample. Table 1 presents statistics of the annotated sample.

3.2.2. Classification schema

In designing our classification schema, we followed multiple criteria. First, we wanted to develop a general-purpose framework which can be easily reused in, adapted to, or extended to other corpora (dialogical or monological). This contrasts with previous studies which proposed highly corpus- or purpose-specific sets of labels (e.g., writer insecurity (Nacey, 2012) or stigmatization (Walker et al., 2025)). This adaptability and transferability are facilitated by the guidelines and by an effort to define clear categories. At the same time, as noted in the literature (Dillon, 1988; Nádraská, 2022), ambiguity and uncertainty are often present in the interpretation of QMUs and it can be difficult to determine the function of a QMU without access to the writer's intention. To address this, our schema allows for the annotation of multiple functions for a single QMU, indicating whether they hold simultaneously or whether they correspond to different plausible but mutually incompatible interpretations (with the addition of an AMBIGUOUS flag).

We also wanted our schema to capture the interpersonal dimension of SQ. During the first stage, we observed that QM frequently appear around words previously used by other participants. In such cases, they often indicate the writer's distancing from the term, functioning as a mechanism of disalignment. We account for this with a specific flag (ECHO).

Our classification is hierarchical, with a high-level distinction between SQ and NSQ. Broadly speaking, SQ express some attitude towards the quoted word or phrase, while NSQ do not. The fine-grained classes presented below clarify what cases fall under each category.[6] Instances can additionally be flagged as UNSURE at both hierarchical levels. More details can be found in the annotation guidelines, and examples are presented in Table 2.

Types of NSQ

- **Meta-linguistic** comments or mentions referring to a word's form, sound or meaning rather than its referent.
- **Direct quotation**. The quoted span reproduces speech, writing, or thought; whether real, invented or hypothetical (Fetzer and Weiss, 2020), in full and in direct speech.
- **Mixed quotation without attitude** (MQwoA). A quoted expression embedded in the writer's own sentence. Although mixed quotation is usually presented as a separate, non-scare function of quotes, Cappelen and Lepore (2003) note the similarities between mixed quoting and SQ, as "reasons for scare quoting are often the same as reasons for mixed quoting." This is why we also consider Mixed quotation *with* attitude under SQ.
- **Emphasis**. Although often regarded as incorrect, QM are sometimes used for emphasis, functioning like italics, boldface or capitalization. Contrary to Dillon (1988), we consider these as NSQ since such usages do not question a word's appropriateness.
- **Designation**. A broad category, not proposed in previous work, encompassing (i) specialized terminology; (ii) names and titles (e.g., movies or abstract ideas); and (iii) objectified expressions used as conceptual or lexicalized units, which may sometimes lead to ungrammaticality if they are not understood as a unit.

Types of SQ

- **SQ Word**. Quoted words whose non-standardness stems from the word form itself (e.g., made-up, incorrect, or so unusual that its correctness is doubted upon). Code-switching

[5]We included all QMUs which were annotated as trigger, indicator or negotiation of some WMN-related phenomenon. 27% of all 2,575 annotated spans in the Reddit portion of NeWMe involved a short QMU.

[6]Although our main interest is on SQ, we adopt a more detailed categorization of NSQ mainly to facilitate their high-level distinction. In practice, it is easier to distinguish the functions of NSQ than of SQ.

	Function	Example
NSQ	META-LINGUISTIC	Take instead the less-controversial word "*desert*."
	DIRECT QUOTATION	Should everybody regardless of their conditions take a happy pill and run around smiling at people and answer "*great!*" if they ask how are you?
	MQwoA	Dawkins defines one gene as "*one replicator*" or "one unit of natural selection" (...)
	EMPHASIS	But Hydropower can change how much goes "*over*" the waterfall versus "under" (via tunnels) (...)
	DESIGNATION (i)	There's actually a term for this - it's called the "*euphimism treadmill*"
	DESIGNATION (ii)	(...) a Christian group which had been very active in trying to prevent it's passage stood outside parliament and recited '*First they Came*' by Martin Niemoller!
	DESIGNATION (iii)	Not that I am subscribing to the "*he drugged her*" theory.
SQ	SQ WORD	This sounds pretty '*science-y*' to me.
	MQwA	Also you are saying that they "*should*" allow you to not see ads. Why should they do that? (...)
	SQ USAGE	Likewise, when trying to "*reason*" with a person who holds an unreasonable position, logic and thoughtful argument doesn't often work.
	SQ USAGE	It consumed ALL of this on an implied social debt to society. It "*borrowed*" from the bank of humanity with the promise of at least trying to repay it.
	SQ USAGE + META-LINGUISTIC	Since "*both sides*" is a fundamentally inaccurate description of US politics, you have to resort to hyperbole to present a balanced picture. But no balance exists.

Table 2: Examples of the functions of quotation marks in our classification schema.

and foreign words would also fall into this category, though none were found in our sample.

- **Mixed quotation with attitude (**MQwA**).** Similar to NSQ's MQwoA, but the quoted words come with a clear evaluative stance or judgment toward what the other person said.
- **SQ Usage**. A general category for SQ not covered by the two labels above. It includes ironic or figurative usages, slang or register mismatches; controversial words or concepts; vague or ambiguous terms (often in a negative light); contextually or situationally inaccurate or inexact words; or instances where the writer distances themselves from the term. We opted against further sub-division of this category due to the difficulty in reliably determining the writer's intention.
- **Echo.** In addition to any SQ label used, this flag marks SQ cases where the quoted expression has already been used in the same conversation, and the writer uses it between quotes to signal disapproval or distancing of another participant's lexical choice.[7]

Multiple annotation. As noted above, multiple functions can be assigned to a single QMU. During the initial annotation stage, we observed recurrent patterns, sometimes combining SQ and NSQ elements. One common case involved critical meta-linguistic comments that question the appropriateness of a word. These were tagged both as META-LINGUISTIC and SQ USAGE, since they exhibit characteristics of both. Other common combinations are discussed in Section 4.1.

[7]Its annotation is facilitated by automatic detection of the quoted passage in previous messages, but human judgment is required to exclude irrelevant matches (e.g., a word used in a different sense) or annotate false negatives which were not caught (e.g., due to typos or morphological variants such as *happy/unhappy*).

3.2.3. Inter-annotator agreement

Once the main annotation was completed, we conducted an inter-annotator agreement (IAA) study on a sample of 250 instances. The second annotator, also an author of this paper, was kept uninformed about the project's early stages to ensure that judgments would rely solely on the annotation guidelines. Both annotators hold a degree in Linguistics and are proficient in English.

The annotator first read the guidelines, after which we discussed any unclear points and introduced the LabelStudio platform. We then entered a training stage consisting of 25 instances. More details about the training and the subsequent modifications made to the guidelines can be found in Appendix D. Finally, the annotator annotated the 250 instances, with access to the updated guidelines but no further feedback. The composition of the sample and IAA results are presented in Section 4.4.

4. Corpus Statistics and Analysis

In this section, we report the main results of our annotation. We start by examining the overall label distribution (Section 4.1) and then focus on the subset of cases involved in WMNs (Section 4.2). We also present a lexical analysis of the quoted passages comparing SQ and NSQ (Section 4.3). Finally, we report the results of the IAA (Section 4.4) and of classification using LLMs (Section 4.5).

Label	Total: freq. (combined)	Total: pct.	Unigrams	ngrams n $\geq$ 1	Single quotes	Double quotes
SQ	1,222	48.9%	53.9%*	41.4%*	51.3%	48.5%
NSQ	916	36.6%	31.7%*	44.0%*	32.9%	37.2%
SQ+NSQ	362	14.5%	14.4%*	14.5%*	15.8%	14.3%
UNSURE	56	2.2%	1.9%	2.8%	2.6%	2.2%
AMBIGUOUS	87	3.4%	3.6%	3.9%	3.5%	3.5%
META-LINGUISTIC	448 (269)	28.7%	7.6%*	21.9%*	27.1%	28.7%
DIRECT QUOTATION	111 (6)	4.7%	2.5%*	8.0%*	1.6%*	5.2%*
MQWOA	22 (20)	1.7%	1.1%*	2.6%*	0.6%	1.9%
EMPHASIS	31 (27)	2.3%	2.9%*	1.5%*	4.2%	2.1%
DESIGNATION	274 (113)	15.5%	7.6%*	27.3%*	16.5%	15.4%
SQ WORD	32 (8)	1.6%	1.8%	1.3%	2.9%	1.4%
MQWA	63 (17)	3.2%	2.9%	4.3%	2.3%	3.6%
SQ USAGE	1,127 (337)	58.6%	63.7%*	50.5%*	62.3%	57.9%
ECHO	0 (599)	23.9%	24.8%*	19.9%*	18.4%	23.5%
UNSURE	0 (39)	1.6%	1.1%	2.4%	1.6%	1.6%
AMBIGUOUS	0 (6)	0.2%	0.1%	0.4%	0.3%	0.2%

Table 3: Frequency of each label. Frequencies in parentheses indicate cases involving a combination of labels (ignoring flag annotations). Percentages are calculated over columns, but using combined frequency counts (e.g., 7.6% of all unigrams had META-LINGUISTIC among their labels). Significant differences are marked with *.

4.1. Global Statistics

Table 3 (first set of columns) presents the results of our annotation. Overall, 48.9% of instances were annotated as SQ, 36.6% as NSQ, and 14.5% as a combination of both. At the fine-grained level, we report how often each label appeared alone versus in combination with other labels, ignoring flags (in parentheses). Among NSQ categories, the most common one is META-LINGUISTIC, followed by DESIGNATION; while EMPHASIS and MQWOA are rare, with fewer than 60 instances each. As expected, most SQ fall under SQ USAGE, whereas SQ WORD occurs only 40 times. The overall low frequency of citation-related categories (DIRECT QUOTATION, MQWA and MQWOA) is expected given our focus on short n-grams. The ECHO flag was used in 599 cases, making up 38% of all SQ and SQ+NSQ instances. This highlights that referring back to an interlocutor's word choices with some display of attitude or distancing is a fairly common phenomenon in this corpus.

A total of 84 instances were additionally marked with UNSURE either at the high- or fine-grained level, meaning that about 3.4% of the data concern tricky cases. The AMBIGUOUS label was also rarely used, particularly at the fine-grained level, likely because multiple interpretations of a QMU were rarely perceived to be mutually exclusive.

Label combinations Combinations were relatively common: 392 instances (16%) received more than one label if ignoring all flags; or 781 (31%) if ECHO is counted as a label. Table 8 in the Appendix shows examples of the seven most common label combinations. Ignoring only the UNSURE and AMBIGUOUS flags, there were 30 different combinations. Most of them (17) combined only two labels. The most frequent pattern, with 368 occurrences, was SQ USAGE + ECHO, showing again the prevalence of this phenomenon. The next most common combination paired the same two labels with a META-LINGUISTIC comment about the word in question (175). The combination of SQ USAGE and META-LINGUISTIC is also quite common without ECHO (52 occurrences). SQ USAGE + DESIGNATION was also frequent (67), sometimes together with ECHO (13), typically when a term was introduced or a category delimited while also questioning its appropriateness. Other combinations were comparatively rare. MQWA and ECHO occur together 19 times, generally when users quoted interlocutors within the same thread. Finally, META-LINGUISTIC + DESIGNATION also occurred 13 times in QMUs introducing and defining a term. All remaining combinations occur 9 times or less.

Results by ngram type The second set of columns in Table 3 compares unigrams to longer n-grams. Chi-square tests of independence ($\alpha = .05$) reveal significant differences for several labels. Unigram QMUs are more often instances of SQ than longer n-grams; however, interestingly, they also exhibit a higher proportion of META-LINGUISTIC cases, as definitions tend to involve single words. Unsurprisingly, DIRECT QUOTATION and DESIGNATION, which often involve phrases, are more common in longer n-grams. This supports the idea that focusing on shorter sequences, especially unigrams, is more effective for capturing instances of SQ.

Results by QM type We examined whether single and double QMs differ in terms of their functions (last columns of Table 3). At the coarse-grained level, they exhibit similar distributions. The analysis at the fine-grained level only reveals one statistically significant difference: double quotes are more often used for DIRECT QUOTATION. This is

Label	WMN		no WMN		Trigger	Indicator	Negotiation
SQ	226	32.8%*	996	55.0%*	54.3%	13.2%	40.1%
NSQ	309	44.8%*	607	33.5%*	28.6%	46.4%	46.5%
SQ+NSQ	155	22.5%*	207	11.4%*	17.1%	40.5%	13.4%
META-LINGUISTIC	232 (148)	55.1%*	216 (121)	18.6%*	39.0%	82.7%	43.3%
DIRECT QUOTATION	28 (3)	4.5%	83 (3)	4.8%	0.0%	1.8%	6.7%
MQwoA	2 (1)	0.4%*	20 (19)	2.2%*	0.0%	0.5%	0.5%
EMPHASIS	7 (8)	2.2%	24 (19)	2.4%	1.9%	1.8%	2.3%
DESIGNATION	33 (13)	6.7%*	241 (100)	18.8%*	4.8%	2.7%	8.3%
SQ WORD	5 (1)	0.9%	27 (7)	1.9%	0.0%	0.5%	1.2%
MQwA	11 (2)	1.9%*	52 (15)	3.7%*	10.5%	1.4%	0.9%
SQ USAGE	210 (152)	52.5%*	917 (185)	60.9%*	61.0%	51.8%	51.4%
ECHO	0 (239)	34.2%*	0 (360)	20.0%*	19.0%	49.1%	31.9%

Table 4: Annotation results on cases involving Word Meaning Negotiation. Percentages are calculated column-wise as in Table 3.

likely related to the fact that single QMs occur more frequently around unigrams (67.7%) than double QMs (58.7%), and unigrams are less often used in DIRECT QUOTATION.[8]

4.2. Results in WMN Sequences

Table 4 presents the results for QMUs occurring within WMNs from NeWMe. Compared to the randomly collected instances, we observe multiple differences: these QMUs involve significantly more NSQ and SQ+NSQ cases. This difference is mainly driven by the fact that they are more often META-LINGUISTIC in nature (55.1% vs 18.2%). QMUs in WMNs are also more likely to bear the ECHO flag, as speakers reuse and contest others' wording.

Breaking the results down by WMN component also reveals interesting paterns. Triggers predominantly involve pure cases of scare quoting. Indicators tend to contain NSQ, but compared to triggers and negotiations, they more often involve QMUs combining characteristics of SQ and NSQ, typically pairing a SQ USAGE with a META-LINGUISTIC comment and ECHO. Negotiations also frequently involve NSQ, although pure SQ are present in about 40% of cases. This confirms that SQ mark problematic usages of words that trigger explicit repair mechanisms, while at the same time a large proportion of the QM used in the subsequent interactions have a non-scare, meta-linguistic component. QMUs in WMNs were also more often annotated with multiple labels (39% compared to 28% of the rest of the corpus). Two combinations stand out: SQ USAGE + ECHO + META-LINGUISTIC (118 cases, 68% of which concern indicators); and SQ USAGE + ECHO (107 cases, 79% of which within a negotiation span). This recurrence suggests that these label combinations may be useful for detecting WMNs involving quotation.

[8] In American English style conventions, single QMs are reserved for nested quotations, whereas the opposite is true of British English. We do not observe consistent adherence to either convention in our corpus.

4.3. Lexical Analysis of Quoted Unigrams

We conducted a series of lexical analyses to compare words appearing in SQ versus NSQ, aiming to better characterize the properties of words used in SQ. Although results are specific to our corpus, this allows us to see how the two types of QMU differ under a similar setting. Statistical significance was tested using Chi-square tests of independence or Mann-Whitney tests, depending on the nature of the data. Full results of the statistical analyses can be found in Appendix C.

Part-of-Speech (PoS) We use spaCy (Honnibal et al., 2020) to PoS-tag posts containing annotated unigram QMUs (774 SQ and 453 NSQ). Chi-square tests revealed that words in SQ were more often verbs, while NSQ contained more proper nouns, pronouns, interjections, subordinate conjunctions and auxiliaries. This intuitively aligns with their functions: verbs tend to have more vague and context-dependent semantics than nouns (Dölling and Heyde-Zybatow, 2007), and speakers may express that through scare quotes. In contrast, proper nouns and interjections are more typically expected in DESIGNATION or DIRECT QUOTATION, such as titles or short expressions ("Yes").

Frequency An analysis by frequency using the `wordfreq` library (Speer, 2022) showed that words in SQ are less frequent than words in NSQ, supporting the idea that SQ involve rarer or less conventional lexical choices.

Number of senses We also compare the number of senses of 686 SQ and 300 NSQ content words using WordNet (Fellbaum, 1998). Mann-Whitney U tests showed that SQ words had significantly more senses on average. This suggests that words in SQ

tend to be more polysemous, and thus potentially more ambiguous, than words in NSQ. This aligns with the finding on PoS, since verbs also tend to have more senses than nouns.

Concreteness We use Brysbaert et al. (2014)'s ratings to compare the concreteness level of 701 SQ and 372 NSQ unigrams. No significant difference was found between the two groups.

Affective properties We use the MPQA Subjectivity Lexicon (Wilson et al., 2005) for subjectivity, the SO-CAL Lexicon (Taboada et al., 2011) for intensity and the NRC Emotion Lexicon (Mohammad and Turney, 2013) for polarity. Based on our observations and the definition of SQ, we expected words in SQ to exhibit higher subjectivity, stronger intensity and a more negative polarity than NSQ words. While our choice of lexica tried to maximize coverage, these experiments are limited by an overall smaller sample size, with a smallest size of 142 NSQ words for intensity. No significant difference was found in subjectivity or intensity. Polarity, however, showed a significant difference in a direction opposite to our expectation: NSQ contained more negative words (31.8%) than SQ (24.3%), while SQ included proportionally more neutral words (42.3% vs 32.7%). It is possible that SQ tend to be used with neutral words and that any perceived negativity or subjectivity stems from the context or the SQ themselves rather than the word. However, this hypothesis would require a more in-depth analysis that we leave for future work.

In summary, our analyses show that words in SQ are more polysemous, less frequent, more likely to be verbs, and more neutral in terms of affect than words in NSQ.

4.4. Inter-Annotator Agreement Results

The composition of the IAA sample is provided in Table 10 in Appendix D. To ensure sufficient representation of all categories, we included at least 14 instances of each fine-grained label and the three most common SQ+NSQ combinations. The remaining instances were randomly sampled to reflect the overall corpus distribution. We calculate Krippendorff's α (Krippendorff, 2013) at the high- and fine-grained levels, both globally and by label. At the high-level, the task was treated as a 3-class problem, but at the fine-grained level, where multiple labels can co-occur, we used an overlap-based distance function, following Garí Soler et al. (2025). Results are presented in Table 5.

Global high-level agreement (.69) fell within the range typically considered reliable ($\alpha \geq .67$) (Krippendorff, 2018). Disagreements mainly concerned mixed SQ+NSQ cases, while pure SQ and NSQ cases reached solid levels of agreement (.77 and .71). Overall agreement at the fine-grained level was lower (.49), which is expected given the higher complexity of the task. Only two labels approached a satisfactory level of α reliability: META-LINGUISTIC and SQ WORD, suggesting these are conceptually clearer and better defined in the guidelines. To contextualize these α values, however, annotators agreed on at least one label in 74% of instances. This indicates that, despite variability in precise label assignment, broad consistency was maintained in the interpretation. Results on combined labels were close to randomness. In fact, global fine-level agreement was much higher on instances with a single label in the reference annotation (.52) compared to those with several labels (.24).

Label	α	Label	α
SQ	.77	SQ USAGE + META	.47
NSQ	.71	SQ USAGE + DESIGNATION	.27
SQ+NSQ	.52	SQ USAGE + EMPHASIS	-.03
SQ WORD	.67	META-LINGUISTIC	.66
MQwA	.27	DIRECT QUOTATION	.64
SQ USAGE	.61	MQwoA	.06
ECHO	.39	EMPHASIS	.38
		DESIGNATION	.39

Table 5: IAA results by label. Global agreement was .69 (high level) and .49 (fine-grained).

We also compared high-level agreement on instances without any UNSURE tag (230 cases) against those where one annotator indicated they were unsure (20 cases). The agreement on the former increased to .74, whereas for the latter it dropped to .07, close to random agreement. This confirms the presence of particularly ambiguous and borderline cases, which can be marked to either be excluded from quantitative analyses or examined in a deeper qualitative analysis.

Overall, results show that the broad SQ/NSQ distinction is reliably identifiable, but the finer categorization is harder to agree upon. Allowing multiple labels allows for richer characterizations but also introduces strong disagreements. Results also point to possible areas of improvement of the guidelines which may benefit from more extensively discussed examples and have subsequently been updated accordingly. Examples of disagreements between the two annotators and a confusion matrix are provided and discussed in Appendix D.

4.5. Automatic Classification with LLMs

To assess the feasibility of performing automatic high-level annotation, we test three open-weight LLMs: OLMo-2-7B-Instruct (Groeneveld et al., 2024), Qwen2.5-7B-Instruct (Qwen-Team, 2024), and Llama-3.3-70B-Instruct. We design 12 different few-shot prompts which differ in their verbosity,

Metric	Qwen	OLMo	Llama	Baseline
α global	.03	-.38	.17	-.27
α SQ	-.03	-.29	.17	-.38
α NSQ	.13	-.16	.27	-.21
α SQ+NSQ	-.02	-.59	.08	-.11
F1 (3-class)	**.52**	.14	.42	.50
F1 (binary)	.75	.59	.77	**.78**
F1 (pure)	.71	.56	**.74**	.73

Table 6: LLM agreement and classification results compared to a majority baseline.

the examples used, and their framing of the task as a 3-class or binary problem. To ensure that model outputs are interpretable, decoding was restricted to the set of expected answer tokens, selecting the one with the highest logit. The best prompts were selected on a development set of 250 instances disjoint from the IAA sample. Full details on the prompts can be found in Appendix E.

For comparison with the second human annotator, we compute Krippendorff's α on the IAA sample. We also evaluate the models on the remaining 2,000 instances. For the 3-class setup, we report micro-averaged F1-scores. In the binary setting, we consider SQ+NSQ as an instance of SQ, as it contains elements of scare quoting; but we also report results excluding these mixed cases ("pure").

Results (Table 6) show overall poor performance. On the IAA sample, several α values reflect randomness and even disagreement with the reference annotation. The worst model overall is OLMo, which strongly over-predicts the SQ+NSQ class, a behavior that is heavily penalized by α. On the test set, performance was almost always worse than a trivial majority baseline that always predicts the most common class in the development set, SQ. In the binary setting, we observe an over-prediction of the SQ class by Llama (72%) and Qwen (83%), when its prevalence in the binarized test set is 64%.

These results indicate that, under few-shot prompting, open-weight LLMs of these sizes struggle with the subtle semantic and pragmatic information required for high-level QMU classification.

5. Discussion and Future Work

We have introduced the first annotated corpus of short quotation mark usages focusing on their interactional functions in conversation. Our annotation schema achieves a reliable agreement at the coarse-grained level, while finer distinctions, especially in multi-label cases, remain more subjective. Our lexical analyses show differences in the words used in scare versus non-scare quotes, for example regarding their part-of-speech and frequency.

Scare quotes are often involved in interactional clashes on word meaning, as reflected by the proportion of WMN spans involving them (15%). They can be triggers of a WMN themselves, and thus problematic word usages, but they are also often observed in indicators (especially when marked with Echo) and negotiations. Crucially, they often signal lexico-semantic misalignment despite apparent superficial lexical alignment. However, a larger proportion of QM in this type of interactions are actually of a non-scare, meta-linguistic nature. While our sampling method does not allow us to estimate the proportion of SQ and meta-linguistic NSQ usages naturally involving WMN, this overlap suggests that these labels could be a useful cue for automatic WMN detection.

An interesting direction for future work, though likely challenging, would be to distinguish between (a) predictable SQ marking an already surprising lexical choice (i.e., an ironic or clearly anomalous usage) and (b) informative SQ that add an interpretive layer to an otherwise natural-sounding usage.[9] De Brabanter (2023) argues that QMs are always optional, but deliberately excludes SQ from the discussion. However, even in cases where QMs are superfluous, their presence helps in processing irony (Schlechtweg and Härtl, 2023). These two cases (predictable and informative SQ) probably exhibit different characteristics and differ in how problematic they are for communication, which may have an impact on their automatic classification. Measures like language model surprisal, which indicates the predictability of a word in a sequence, could potentially help capture this distinction.

We also plan to extend our annotation to cover other genres, situations, speakers and periods, which will allow us to test the robustness of our annotation schema; and eventually propose a finer-grained taxonomy of SQ types. Another direction will be to test supervised models trained on this or extended data, which may outperform LLMs and eventually support large-scale SQ prediction and analysis; for example, to investigate claims that their frequency has increased (McCullagh, 2017). Detecting scare quotes could benefit NLP tasks where understanding subtle speaker intentions is crucial: properly interpreting sarcasm or criticism can affect the outputs of machine translation, summarization, sentiment analysis or dialogue systems.

Overall, we have seen that scare quotes play an important role in managing meaning in interaction. Our annotations, which are made publicly available, will facilitate further empirical research on scare quoting.

[9] For example, *The man eating the cake is on a "diet"* and *I really like this "painting"*.

6. Limitations

The main limitation of our study is the use of a single corpus with very specific characteristics (i.e., debate-like exchanges). However, although this limits the usefulness of our new resource, we still see our work as a valuable first step. The use of scare quotes likely differs across genres and conversational situations; for this reason we do not generalize our findings beyond the dataset studied and plan to extend the annotation to a broader spectrum of interactions in future work. Our analysis also focuses exclusively on English, but speakers may use scare quotes differently in other languages, both functionally and in terms of punctuation conventions.

We also acknowledge the potential subjective bias associated with having a single main annotator, as well as the possible bias arising from the non-fully-random selection of the sample. Involving additional annotators, conducting agreement on a larger subset, and implementing a full double annotation would strengthen the resource and should be considered in future work. This would be particularly relevant for fine-grained categories, where distinctions are more challenging. At the same time, our two rounds of guidelines refinement provide a foundation for future efforts in this direction.

Additionally, the fact that our lexical analyses are limited to unigrams excludes from our conclusions a portion of quoted items that could provide a more complete picture of scare quote usage. However, they would require a dedicated methodology and a more qualitative analysis, which was beyond the scope of this work.

7. Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was partially funded by the Agence Nationale de la Recherche SINNet project (ANR-23-CE23-0033-01). Additional support was provided by the ANR under the France 2030 program PRAIRIE (ANR-23-IACL-0008).

8. Bibliographical References

Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee. 2021. Testimonial injustice: linguistic bias in the medical records of black patients and women. *Journal of general internal medicine*, 36(6):1708–1714.

Herman Cappelen and Ernie Lepore. 1997. Varieties of quotation. *Mind*, 106(423):429–450.

Herman Cappelen and Ernie Lepore. 2003. Varieties of quotation revisited. *Belgian Journal of Linguistics*, 17(1):51–75.

Letizia Cirillo. 2019. The pragmatics of air quotes in english academic presentations. *Journal of Pragmatics*, 142:1–15.

Philippe De Brabanter. 2010. The semantics and pragmatics of hybrid quotations. *Language and Linguistics Compass*, 4(2):107–120.

Philippe De Brabanter. 2017. Why quotation is not a semantic phenomenon, and why it calls for a pragmatic theory. In *Semantics and pragmatics: Drawing a line*, pages 227–254. Springer.

Philippe De Brabanter. 2023. Quotation does not need marks of quotation. *Linguistics*, 61(2):285–316.

George Dillon. 1988. My words of an other. *College English*, 50(1):63–73.

Johannes Dölling and Tatjana Heyde-Zybatow. 2007. Verb meaning: How much semantics is in the lexicon. *Interface and interface Conditions*, pages 33–76.

Anita Fetzer and Daniel Weiss. 2020. Doing things with quotes: Introduction. *Journal of Pragmatics*, 157:84–88.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring Lexico-Semantic Alignment in Debates with Contextualized Word Representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 50–63, Toronto, Canada. Association for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2025a. Potentially Problematic Word Usages and How to Detect Them: A Survey. *Accepted at the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024.

OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Gutzmann and Erik Stei. 2011. How quotation marks what people do with words. *Journal of Pragmatics*, 43(10):2650–2663.

Lauri Haapanen and Daniel Perrin. 2017. Media and quoting: Understanding the purposes, roles, and processes of quoting in mass and social media. In *The Routledge handbook of language and media*, pages 424–441. Routledge.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ken Hyland. 2010. Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies*, 9(S2):125–143.

K. Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Agnieszka M Kula and Monika Grzelka. 2022. Quotation in Social Media: How Sharing Other People's Words Could Increase Misinformation. *Studia Humanistyczne AGH*, 21(2):81–98.

Tom McArthur, Jacqueline Lam-McArthur, and Lise Fontaine. 1992. The Oxford companion to the English language. *Oxford: Oxford University Press*.

Mark McCullagh. 2017. Scare-quoting and incorporation. In *The semantics and pragmatics of quotation*, pages 3–34. Springer.

Andreas Musolff. 2015. Quotation and online identity: The voice of tacitus in german newspapers and internet discussions. *The Pragmatics of Quoting Now and Then*, 89:125.

Susan Nacey. 2012. Scare quotes in Norwegian L2 English and British English. *English Corpus Linguistics: Looking back, Moving forward*, page 117.

Zuzana Nádraská. 2022. The function of scare quotes in hard news: metadiscoursal and generic perspectives. *Discourse and Interaction*, 15(2):101–127.

Anna P. Goddu, Katie J O'Conor, Sophie Lanzkron, Mustapha O Saheed, Somnath Saha, Monica E Peek, Carlton Haywood Jr, and Mary Catherine Beach. 2018. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33(5):685–691.

Sean Papay and Sebastian Padó. 2019. Quotation detection and classification with a corpus-agnostic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 888–894, Varna, Bulgaria. INCOMA Ltd.

Gina M Piscitello, Ruthe Ali, Katrina Hauschildt, and Jane Schell. 2025. How Do Clinicians Use Quotations in Goals of Care Notes? *Chest*.

Stefano Predelli. 2003. Scare quotes and their relation to other semantic issues. *Linguistics and philosophy*, 26(1):1–28.

Qwen-Team. 2024. Qwen2.5: A party of foundation models.

François Recanati. 2001. Open quotation. *Mind*, 110(439):637–687.

Claudia Ricci and Corinne Rossari. 2018. Commitment phenomena through the study of evidential markers in romance languages.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics.

Marcel Schlechtweg and Holden Härtl. 2020. Do we pronounce quotation? an analysis of name-informing and non-name-informing contexts. *Language and Speech*, 63(4):769–798.

Marcel Schlechtweg and Holden Härtl. 2023. Quotation marks and the processing of irony in English: evidence from a reading time study. *Linguistics*, 61(2):355–390.

Barbara Schneider. 2002. Nonstandard quotes: Superimpositions and cultural maps. *College Composition & Communication*, 54(2):188–207.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

University of Chicago Press. 2017. The chicago manual of style online. https://www.chicagomanualofstyle.org/. 17th edition. Accessed 2025-08-13.

Andrew Walker, Annie Thorne, Sudeshna Das, Jennifer Love, Hannah LF Cooper, Melvin Livingston, and Abeed Sarker. 2025. CARE-SD: classifier-based analysis for recognizing provider stigmatizing and doubt marker labels in electronic health records: model development and validation. *Journal of the American Medical Informatics Association*, 32(2):365–374.

Bingjuan Xiong and Jessica S Robles. 2023. Functions of quotation in online political comments. *Discourse, Context & Media*, 55:100717.

9. Language Resource References

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Aina Garí Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. 2025. The NeWMe Corpus: A gold standard corpus for the study of Word Meaning Negotiation. In *PREPRINT (Version 1) available at Research Square. To appear in Language Resources and Evaluation.*

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

A. Examples of Word Meaning Negotiation

Word Meaning Negotiation (WMN) sequences can reflect a problem of misunderstanding, non-understanding or disagreement with regard to word meaning. Sequences in this debate-centered Reddit corpus often involve disagreement. See Table 7 for examples of WMN involving quotation marks.

B. Annotation Results: Label Combinations

Table 8 contains the frequencies and examples of the six most common label combinations found in our annotation.

C. Full Lexical Analysis Results

This section provides additional details on the lexical analyses conducted on quoted unigrams.

PoS We analyze 774 SQ and 453 NSQ unigrams. Chi-square tests of independence were run for every PoS occurring at least 3 times in one subset. Table 9 summarizes the main results.

Frequency We apply Mann-Whitney tests on "zipf frequencies" (base-10 logarithm of occurrences per billion words, 774 SQ and 453 NSQ). SQ unigrams are significantly less frequent than NSQ ($p = .004$, $U = 158245.5$, medians: 4.56 vs 4.73).

Number of senses The number of senses is calculated aggregating all possible senses of the unigram regardless of PoS (686 SQ and 300 NSQ). SQ words exhibit significantly more senses: $p < .001$, $U = 116552.0$, mean values 8.3 and 6.5.

Concreteness Mann-Whitney tests to analyze 701 SQ and 372 NSQ unigrams for concreteness did not show any significant difference between SQ and NSQ ($p = .585$, $U = 133027.5$).

Subjectivity and intensity The subjectivity analysis included 327 SQ and 168 NSQ unigrams which were classified as having a weak or strong subjectivity. Chi-square tests revealed no statistically significant difference ($p = .057$). 45.9% of unigrams in SQ were strongly subjective, against 44.6% in NSQ. To derive intensity scores from SO-CAL, we took the absolute value of the score associated with every word. For 301 SQ and 142 NSQ unigrams, Mann-Whitney tests showed no significant difference ($p = .172$, $U = 19730.0$).

A: (...) If the only way that a person can get “married” is to do so in a way that gives none of the rights that other groups of people who are married get is exactly equal. Their marriages are void. They aren’t married (...)
B: In my opinion, that isn’t what marriage is about. Filling a box saing that you are married isn’t what marriage is about. It is an emotional attachment to another person. No state can truly ever take that away from two people. (...)
C: in the u.s., a marriage is a legal contract between spouses that establishes rights and obligations between them, between them and their children, and between them and their in-laws. on a governmental standpoint, it has **nothing** to do with love.

A: If you call some people "normal" then you are calling the others "abnormal". And the word "abnormal" has a negative connotation (...)
B: Regardless of connotations or how people feel about words, abnormal does not mean negative. (...)
A: (...) If you tell me a word, it dosen’t matter what its dictionary definition is. The only things that do matter are what you wanted to convey with that word and how I understood it. When you call something "normal" other people hear you saying "this is the way it should be". If "normal" really only meant "most common" then there would be no controversy and this thread wouldn’t exist.

A: It’s not my job to take care of drunk friends (...) it’s not acceptable that people just drink too much and expect their friends to take care of them.
B: I think you need to seriously reevaluate your definition of the word "friend".
(*Negotiation missing*)

Table 7: Examples of annotated Word Meaning Negotiation sequences from NeWMe (Garí Soler et al., 2025) and Winning Arguments (Tan et al., 2016) involving quotation marks. Triggers are marked in green, indicators in red and negotiations in blue.

Combination	Freq.	Example
SQ Usage + Echo	368	[STA-CITE] Photographers were reviled by painters, digital photographers reviled by film photographers, photoshoppers reviled by “pure digital photographers,” and so on and so forth. [END-CITE]I’m not “reviling” digital cinematography at all.
SQ Usage + Meta-linguistic + Echo	175	This quote could be held as true only under the premise that “religion” in this case does not refer exclusively to theistic religions. Any system of rituals or beliefs that people follow, including the patriotic following of orders you referred to, is a religion in this context.
SQ Usage + Designation	67	Honestly I think you should change your view on the basis that a framework for mitigating the impact on debt repayment by the poor already exists and that simply expanding those protections will achieve much of what you want without requiring a “debt jubilee” which would be economically disruptive.
SQ Usage + Meta-linguistic	52	Most people use the word “humbled” to mean the opposite of what it actually means, and should stop immediately.
MQwA + Echo	19	I think you’ve got two separate issues here. One is what you posted about, which is whether these days should be “considered” Federal holidays, and whether they “warrant” a company deciding to close and give everyone the day off.
Meta-linguistic + Designation	13	The name “Canada” was also derived from an Iroquois word, so “Native Canadians”, while also a pretty outmoded term these days, neatly avoids exactly the issue OP was worried about anyway.
SQ Usage + Echo + Designation	13	Premise 2 is true because time literally began with the big bang; thus, there is no “before” the big bang.

Table 8: Frequency and examples of the six most common label combinations. Target QMUs are underlined and in blue.

PoS	p-value	χ^2	% in SQ	% in NSQ
VERB	<.001	33.86	18.9	6.6
PROPN	<.001	27.93	1.6	7.7
INTJ	<.001	24.72	0.4	4.6
PRON	<.001	14.77	0.9	4.4
SCONJ	<.001	14.60	0.0	2.2
NUM	.014	6.07	0.0	1.1
AUX	.016	5.82	0.5	2.2
ADJ	.102	2.67	35.7	30.6
NOUN	.186	1.75	39.5	35.3
ADV	.433	0.61	2.2	3.1

Table 9: Results of the PoS analysis.

Polarity Our analysis relied on 428 SQ vs 211 NSQ unigrams. Chi-square tests showed a significant difference (p = .044, χ^2= 6.26). SQ consisted of 42.3% neutral, 33.2% positive and 24.3% negative words, while NSQ included 32.7% neutral, 33.6% positive and 31.8% negative words. Small percentages not accounted for correspond to words belonging to both polarities.

D. IAA Details

D.1. Training stage

The training sample consisted of 25 instances (12 SQ, 9 NSQ, and 4 SQ+NSQ cases). After reviewing it, disagreements were discussed in a feedback session. As a result, a few modifications were made to the guidelines, notably making a clearer distinction between the ECHO and MQwA labels, and emphasizing the possibility – and desirability – of assigning multiple labels to an instance, instead of trying to choose the most representative one. While some systematic disagreements could be detected and corrected, other differences, stemming from different interpretations of a situation, are expected and unavoidable.

D.2. IAA sample composition

Table 10 contains information on the labels of the 250 instances used in the inter-annotator agreement study. As explained in the paper, we included at least 14 instances of each fine-grained label and the three most common SQ+NSQ combinations. The remaining instances were randomly sampled to reflect the overall corpus distribution.

D.3. Examples of disagreement

Figure 1 shows the confusion matrix for the high-level labels. The most notable disparity concerns NSQ and SQ+NSQ: instances classified as SQ+NSQ by the main annotator were sometimes labeled as NSQ by the second annotator. A closer examination of the corresponding fine-grained labels reveals that, in 8 of the 14 cases, the disagreement involves the label EMPHASIS, which the main annotator treated as co-occurring with SQ USAGE.

In Table 11, we present three examples of disagreements between the two annotators. In the first example, one annotator saw "social contract" as neutrally used terminology and annotated it as DESIGNATION, while the other saw this particular instance of "social contract" in quotes as expressing the non-literal, abstract nature of the term. In the second example, the main annotator saw this again as an instance of SQ USAGE, with the speaker expressing that their word "action" may not have been a clear lexical choice. The second annotator saw this instead as an instance of MQwoA, as the speaker is citing themselves and simply clarifying what they meant. Finally, in the third example, both annotators agreed that "inherently" could be an echoed SQ USAGE, but the main annotator also saw this as a potentially AMBIGUOUS case with EMPHASIS.

These cases serve to demonstrate the subjectivity of the task and how the same situation can be interpreted in multiple valid ways.

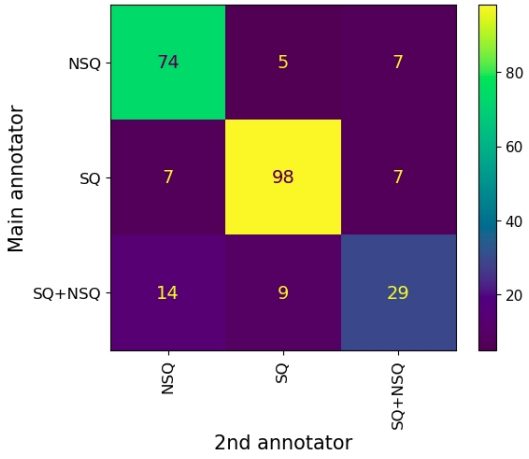

Figure 1: Confusion matrix between the reference annotator and the 2nd annotator for high-level distinctions.

Label	Freq.	Pct.
SQ	112	44.8%
NSQ	86	34.4%
SQ+NSQ	52	20.8%
META-LINGUISTIC	25	10%
DIRECT QUOTATION	17	6.8%
MQWOA	14	5.6%
EMPHASIS	15	6.0%
DESIGNATION	14	10.0%
SQ WORD	15	6.0%
MQWA	17	6.8%
SQ USAGE	80	32.0%
SQ USAGE + META-LINGUISTIC	19	7.6%
SQ USAGE + DESIGNATION	15	7.2%
SQ USAGE + EMPHASIS	15	6.0%
Other combinations	4	1.6%

Table 10: Composition of the IAA sample.

E. LLM prompts

Figures 2 and 3 present the prompt templates used for the 3-class and the binary settings respectively. The class definitions with three different levels of verbosity can be found in Figures 4 and 5. Finally, the sets of examples used are presented in Figures 6 and 7. The Llama and Qwen models worked best with the prompts with the lowest verbosity level and the smaller set of examples (set B) in both the 3- and 2-class settings. OLMo, instead, obtained the best results with the medium verbosity level, using example set A in the 3-class setting and example set B in the binary setting. Llama-3.3-70B-Instruct was used with 4bit quantization.

Reference	2nd annotator	Utterance
SQ Usage	Designation	I would like to talk about this statement you keep making: [STA-CITE]>but consensus can't make something true [END-CITE]I think this doesnt hold up in certain models of morality. There is a certain model called the [Social contract](http://en.wikipedia.org/wiki/Social_contract) perhaps most famously postulated by Thomas Hobbes (although I think it came from Epicurus first), it states (in regards to ethics) that we only abide by ethics because then in turn, other people will abide by the same ethics, ergo in this model, what everyone does actually will dictate what is moral, after all, you are all in the 'social contract'. Ethical discussion do have a habit of boiling down to 'your model of ethics versus mine', but I think this is an example in which ethics are indeed relative due to society's views. So by a technicality your title is not sound.
SQ usage	MQwoA	The "action" I'm talking about is not necessarily going to war with these people. I make donations to charities that help at refugees for example. These videos will exist for people to watch regardless of who watches them, or who doesn't. Just because they can be used for recruitment does not mean they cannot be used to show their crimes.
SQ Usage + Echo + Emphasis	SQ Usage + Echo	[STA-CITE]> None of these things are inherently cheap - it's just that oftentimes, when a director pulls one of these out, he/she uses it cheaply. [END-CITE]They are indeed not 'inherently' cheap, just like a prince saves the princess is not an 'inherently' cheap plot device. But I think both became cheap by being overdone and becoming a simple addition in a story to transform in something that is intended.

Table 11: Examples of disagreements between annotators. The target QMU is underlined and in blue.

LLM Prompt Template (3 classes)

- Task: Classify the use of quotation marks into scare-quotes (S) or non-scare quotes (N), or Both (B).
- Definitions:

{DEFINITIONS}

- Rules:

1. Classify only the text inside <target>...</target>.
2. Answer with exactly ONE of: S / N / B.

- Examples:

{EXAMPLES}

- Your turn:

Utterance: {UTTERANCE}
Your answer (S, N, or B):

Figure 2: Template of the prompts shown to LLMs for the 3-class setting.

LLM Prompt Template (2 classes)

- Task: Classify the use of quotation marks into scare-quotes (S) or non-scare quotes (N).
- Definitions:

{DEFINITIONS}

- Rules:

1. Classify only the text inside <target>...</target>.
2. Answer with exactly ONE of: S / N.

- Examples:

{EXAMPLES}

- Your turn:

Utterance: {UTTERANCE}
Your answer (S or N):

Figure 3: Template of the prompts shown to LLMs for the 2-class setting.

Definitions (3 classes)

High verbosity

- S (Scare Quotes): The expression in quotes is not to be taken at face value. For example, the expression is a suboptimal lexical choice (e.g., because it doesn't fully apply to the situation being described, or because the writer is unsure that this is the correct word to use), it is used in a figurative sense, or with irony (e.g., "the 'generous' lady donated one cent to charity"). Alternatively, the writer seems to want to distance themselves from the usage, perhaps because the word is connotatively loaded or controversial, or because it comes from a different register (e.g., "This lady works as an 'influencer' on Instagram"). It can also be a word that is unusual, incorrect or made-up. Finally, the writer may express, implicitly or explicitly, some attitude towards someone else's words.
- N (Non-Scare Quotes): The expression in quotes is to be taken at face value, with no added attitude. For example, quotes enclose a meta-linguistic usage ("'Cat' has three letters"), they are used to cite someone else ("He said 'I'm hungry'"), or mark specialized terminology, titles ("Have you watched 'Gone with the wind?'") or abstract categories. They can also be used for emphasis.
- B (Both): Cases where content between quotes exhibits characteristics from both S and N. For example, a meta-linguistic comment about word meaning accompanied by some distancing or criticism.

Medium verbosity

- S (Scare Quotes): The expression in quotes is not to be taken at face value. For example, the expression is a suboptimal lexical choice, it is used in a figurative sense, or with irony. Alternatively, the writer seems to want to distance themselves from the usage because the word is connotatively loaded, controversial, or because it comes from a different register. It can also be a word that is unusual, incorrect or made-up. Finally, the writer may express, implicitly or explicitly, some attitude towards someone else's words.
- N (Non-Scare Quotes): The expression in quotes is to be taken at face value, with no added attitude. For example, quotes enclose a meta-linguistic usage, they are used to cite someone else, or mark specialized terminology, titles or abstract categories. They can also be used for emphasis.
- B (Both): Cases where content between quotes exhibits characteristics from both S and N. For example, a meta-linguistic comment about word meaning accompanied by some distancing or criticism.

Low verbosity

- S (Scare Quotes): The expression in quotes is not to be taken at face value (suboptimal lexical choices, figurative or ironic usages, controversial concepts, incorrect or made-up words...) or some attitude is expressed about it.
- N (Non-Scare Quotes): The expression in quotes is to be taken at face value (meta-linguistic usages, citations, specialized terminology, titles, abstract categories or emphasis).
- B (Both): Cases where content between quotes exhibits characteristics from both S and N. For example, a meta-linguistic comment about word meaning accompanied by some distancing or criticism.

Figure 4: Definitions of each class at different levels of verbosity shown to LLMs in the 3-class setting.

Definitions (two classes)

High verbosity

- N (Non-Scare Quotes): The expression in quotes is to be taken at face value, with no added attitude. For example, quotes enclose a meta-linguistic usage ("'Cat' has three letters"), they are used to cite someone else ("He said 'I'm hungry'"), or mark specialized terminology, titles ("Have you watched 'Gone with the wind?'") or abstract categories. They can also be used for emphasis.
- S (Scare Quotes): The expression in quotes is not to be taken at face value. For example, the expression is a suboptimal lexical choice (e.g., because it doesn't fully apply to the situation being described, or because the writer is unsure that this is the correct word to use), it is used in a figurative sense, or with irony (e.g., "the 'generous' lady donated one cent to charity"). Alternatively, the writer seems to want to distance themselves from the usage, perhaps because the word is connotatively loaded or controversial, or because it comes from a different register (e.g., "This lady works as an 'influencer' on Instagram"). It can also be a word that is unusual, incorrect or made-up. Finally, the writer may express, implicitly or explicitly, some attitude towards someone else's words, even if there is some meta-linguistic comment about them.

Medium verbosity

- N (Non-Scare Quotes): The expression in quotes is to be taken at face value, with no added attitude. For example, quotes enclose a meta-linguistic usage, they are used to cite someone else, or mark specialized terminology, titles or abstract categories. They can also be used for emphasis.
- S (Scare Quotes): The expression in quotes is not to be taken at face value. For example, the expression is a suboptimal lexical choice, it is used in a figurative sense, or with irony. Alternatively, the writer seems to want to distance themselves from the usage because the word is connotatively loaded, controversial, or because it comes from a different register. It can also be a word that is unusual, incorrect or made-up. Finally, the writer may express, implicitly or explicitly, some attitude towards someone else's words, even if there is some meta-linguistic comment about them.

Low verbosity

- N (Non-Scare Quotes): The expression in quotes is to be taken at face value (meta-linguistic usages, citations, specialized terminology, titles, abstract categories or emphasis).
- S (Scare Quotes): The expression in quotes is not to be taken at face value (suboptimal lexical choices, figurative or ironic usages, controversial concepts, incorrect or made-up words...) or some attitude is expressed about it, even if there is some meta-linguistic comment about them.

Figure 5: Definitions of each class at different levels of verbosity shown to LLMs in the 2-class setting.

Few-shot Examples (three classes)

Set A
Utterance: I do not know the meaning of <target>'diplomacy'</target>.
Answer: N
Utterance: This so-called <target>'diet'</target> allows him to eat ice-cream.
Answer: S
Utterance: <target>'Masculine'</target> is a wrong way of describing this person.
Answer: B
Utterance: He said <target>'hello'</target> and went to grab some coffee.
Answer: N
Utterance: He said he was the <target>'humblest person ever'</target>, can you believe that?
Answer: S
Utterance: I think you don't know what a <target>'promise'</target> means, it is not about trying to do something, it is about actually doing it.
Answer: B
Set B
Utterance: What does <target>'errand'</target> mean?
Answer: N
Utterance: Phagocytosis means that, we could say, cells <target>'eat'</target> other cells.
Answer: S
Utterance: <target>'Moral'</target> is too vague of a concept, you need to speak clearly.
Answer: B

Figure 6: Sets of examples provided to the LLMs in the 3-class setting.

Few-shot Examples (two classes)

set A
Utterance: I do not know the meaning of <target>'diplomacy'</target>.
Answer: N
Utterance: This so-called <target>'diet'</target> allows him to eat ice-cream.
Answer: S
Utterance: He said <target>'hello'</target> and went to grab some coffee.
Answer: N
Utterance: He said he was the <target>'humblest person ever'</target>, can you believe that?
Answer: S
Set B
Utterance: What does <target>'errand'</target> mean?
Answer: N
Utterance: Phagocytosis means that, we could say, cells <target>'eat'</target> other cells.
Answer: S

Figure 7: Sets of examples provided to the LLMs in the 2-class setting.