Language Resources and Evaluation (2026) 60:38
https://doi.org/10.1007/s10579-026-09907-x

ORIGINAL PAPER

The NeWMe corpus: a gold standard corpus for the study of word meaning negotiation

Aina Garí Soler[1,2] · Jenny Myrendal[3,4] · Chloé Clavel[1,2] · Staffan Larsson[3]

Received: 6 February 2025 / Accepted: 12 February 2026

Abstract

Word Meaning Negotiation (WMN) sequences occur when participants focus on clarifying or negotiating the meaning of a word or phrase, often prompted by questions or challenges. These interactions temporarily shift the conversation to explore nuances of meaning—sometimes resulting in quick clarification when due to insufficient understanding of word meaning, and other times leading to extended debates, such as disagreements on what a word can or should mean. This paper presents the largest, freely available manually annotated corpus of WMNs to date, encompassing spoken dyadic and multiparty conversations as well as online discussions. Our methodology combines searching for WMNs using regular expressions with a detailed annotation scheme that categorizes WMNs into types triggered by non-understanding (NONs: Non-understanding WMN) or disagreement (DINs: Disagreement WMN), and distinguishes between negotiations of situated and potential meanings. We also annotate incomplete negotiations and related phenomena, and analyze inter-annotator agreement to evaluate the reliability of the annotation schema. Preliminary investigations of WMNs in the corpus reveal distinct patterns in WMNs across contexts, with NONs prevalent in spoken interactions and DINs dominating online debates. This resource lays a foundation for studying semantic alignment, developing automated WMN detection, and creating adaptive dialogue systems. Our findings highlight the complexity of WMNs and provide practical insights for their identification and analysis.

Keywords Word meaning negotiation · WMN · Semantic alignment · Semantic coordination · Interactional linguistics · Misunderstanding

Aina Garí Soler and Jenny Myrendal contributed equally to this work.

Extended author information available on the last page of the article

Published online: 10 April 2026

1 Introduction

Mutual understanding is a fundamental aspect of successful interaction, enabling participants to align their interpretations and ensure that communication flows freely without major interruptions. In any conversation, alignment between interlocutors can occur at various levels, including lexical choice (Brennan & Clark, 1996), syntax (Branigan et al., 2000) and conceptual representations (Stolk et al., 2016).[1] Alignment at the level of word meaning, which is the focus of this study, is both essential and sometimes challenging to achieve. When participants share a common understanding of word meanings, conversations tend to proceed more smoothly.

As Garfinkel (1967) observed with the concept of "current practical purposes," interactions typically require just enough alignment to continue, even without full mutual understanding. Research has shown that achieving alignment at one level can support alignment at other levels (Cleland & Pickering, 2003; Pickering & Garrod, 2004). However, conceptual misalignments can sometimes remain hidden beneath apparent lexical alignment, and, if overlooked, have the potential to cause significant misunderstandings (Garí Soler et al., 2023; Schober, 2005). Despite its importance, conceptual (mis)alignment remains a relatively underexplored area, primarily due to the challenges involved in identifying it.

When misalignments are detected, they may lead to explicit efforts to clarify or resolve the issue. Such efforts often involve meta-linguistic discussions, where participants directly address and negotiate the meanings or uses of problematic words or phrases. These interactions, which we call Word Meaning Negotiation (WMN; Myrendal, 2015), play an important role in achieving alignment. For instance, if one speaker uses the word X and another speaker questions whether X is appropriate in the current context, the conversation may temporarily shift from the main topic to focusing on language, specifically the meaning of X. Through such sequences, participants engage in a negotiation process to explore and address the meaning of the word. They may contribute their individual perspectives, propose what the word should mean in the given context, or challenge existing interpretations by highlighting what the word cannot mean. This exchange allows participants to collaboratively explore and shape the boundaries and potentials of the word's meaning, whether the interaction involves addressing disagreement or clarifying its meaning in context.

The term *negotiate* is used broadly to encompass a range of actions and contributions, including explaining, problematizing, questioning, or exemplifying word meanings. Similarly, the term *word* is used here in an inclusive sense, referring not only to single word forms but also to lexicalized expressions or phrases. This broader definition recognizes that meaning negotiation can involve both individual lexical items and multi-word units that function as cohesive semantic entities in interaction.

Consider a scenario where one participant describes a colleague as *ambitious* and another responds, "What do you mean by ambitious: driven or just self-centered?"

[1] We take lexical alignment to be an agreement on what word to use for a given meaning, and conceptual alignment to be an agreement about what meaning to assign a given word. While conceptual alignment is the prototypical case of a WMN, lexical alignment also involves WMN since when deciding on which word to use for a meaning, one is also deciding on the meaning of that word (and the alternative).

The dialogue may then unfold with each participant elaborating on their understanding of the word *ambitious* in the given context, aiming to achieve enough common ground to move on in the conversation. This sequence illustrates a typical WMN, where the meaning of a word becomes the focus of collaborative negotiation, temporarily shifting the conversation's focus to language itself.

Word Meaning Negotiation is important for fostering conceptual and linguistic alignment, particularly in contexts where achieving mutual understanding is more complex. Misunderstandings arising from conceptual misalignments can disrupt communication, and WMNs serve as a mechanism to identify and resolve these issues. By explicitly addressing word meanings, WMNs facilitate collaborative efforts to clarify, refine, or redefine meanings, ensuring the conversation can proceed smoothly.

The importance of WMN may be especially relevant in environments where traditional cues for achieving alignment are limited or absent. For instance, in online interactions, participants often lack access to nonverbal cues such as gestures, gaze, and intonation, which support alignment in face-to-face conversations (Babcock et al., 2014). This limitation increases the likelihood of misalignments and highlights the need for WMN as a tool to navigate such challenges.

While WMN resolves immediate misalignments in conversation, studying these processes has broader implications for understanding interactional dynamics. It offers insights into how participants collaboratively shape language and negotiate meaning in real time as part of everyday activities and interactions. By studying WMN, we can better understand the mechanisms that underpin alignment at both lexical and conceptual levels, as well as explore its potential to address controversial or ambiguous terms.

This paper outlines the process of establishing a gold standard corpus for the detection and study of WMN sequences, presenting what we believe to be the first manually annotated corpus of its kind, encompassing both spoken and written conversations. We propose and apply a methodology to locate likely WMN sequences, and provide an annotation schema with detailed instructions for the annotation of WMN and related phenomena observed in the data. In addition to the primary WMN types we focus on later, our analysis identified several related phenomena that we describe in detail in Sect. 4.1.2. The corpus, which will be made publicly available, includes expert annotations and we describe the process of training multiple annotators and calculating inter-annotator agreement (IAA).

While the study does not directly explore WMN as a mechanism of interactional alignment, it builds on qualitative studies that have examined the phenomenon qualitatively in detail. These studies have provided insights into the characteristics of WMN, its functions, and its implications for human interaction, exploring its role in conversational repair (Myrendal, 2025), disagreements (Myrendal, 2019), and semantic conflicts about moral issues (Myrendal & Larsson, 2025). However, the fact that a corpus of WMNs has been lacking means that results so far have been mostly qualitative. In our study, we identify WMNs in 11% of the conversations considered for annotation. While our method does not allow us to determine their actual prevalence, this proportion suggests that WMNs are challenging to detect but sufficiently frequent to deserve further investigation.

In addition to the theoretical contributions of the annotation schema and instructions, the resulting annotations facilitate future quantitative and qualitative investigation of WMN. These annotations can also be used to identify potentially controversial or problematic terms, and to investigate the mechanisms of conceptual or lexico-semantic alignment. Together, this foundational corpus and previous qualitative analyses contribute to both the theoretical understanding and practical applications of WMN, for example in developing automated dialogue systems capable of detecting, negotiating and adapting to unexpected, problematic or controversial word usages.

The paper is organized as follows: Sect. 2 reviews previous work on WMN and related concepts, and is followed by a detailed definition of WMN in Sect. 3. In Sect. 4, we outline the corpora that form the raw materials for the NeWMe corpus and explain our annotation procedure. Section 5 presents an analysis of our annotations, and Sect. 6 describes the inter-annotator agreement study. Finally, we conclude and discuss directions for future work in Sect. 7.

2 Related work

Repair and clarification are fundamental mechanisms for managing communication breakdowns, and they have been studied extensively in both Second Language Acquisition (SLA) and Conversation Analysis (CA). While SLA often focuses on repair in learning contexts, CA examines repair as a broader phenomenon in natural interaction.

Repair involves interrupting the conversational flow to resolve issues related to speaking, hearing, or understanding. A frequent feature of dialogue, repairs occur roughly once every 25 words in conversational speech (Hough & Purver, 2013). They play a crucial role in ensuring the smooth progression of interaction by maintaining or restoring intersubjectivity (Schegloff, 2007). Repairs can target a range of problems, from lexical ambiguities, such as the use of a problematic word, to more general issues like mishearing or disruptions in turn-taking. WMN sequences can be regarded as a type of conversational repair specifically targeting word meaning.

Clarification requests, a key form of other-initiated repair, play an important role in addressing breakdowns in understanding and managing interactional problems. As Purver (2004) highlights, clarification requests vary widely in both their form and the type of information they seek to resolve. They may take explicit forms, such as detailed queries (e.g., “What do you mean by X?”) or minimal and ambiguous expressions like “What?” that rely on context and intonation for interpretation. These diverse forms reflect the flexibility of clarification requests in targeting different aspects of an utterance, from the meaning of specific words to their appropriateness in context.

Purver’s taxonomy identifies explicit forms that clearly specify the issue and minimal forms that rely on context and intonation to clarify ambiguities. For instance, explicit queries like “What do you mean by X?” demand detailed responses, while minimal forms such as “X?” or “What?” signal a general difficulty with understanding the prior turn without pinpointing a specific trouble source. Drew (1997) explores these minimal repair initiators, referred to as “open-class repair initiators”, which

do not locate a specific repairable but instead indicate a broader problem with the prior utterance. These initiators are particularly relevant when a speaker signals an issue in understanding the sequential connection of an utterance rather than a specific lexical ambiguity. Drew's analysis highlights how open-class repair initiators are frequently used when speakers have difficulty grasping why something was said or how it relates to the prior discourse, underlining the connection between repair and alignment in conversation.

Research within the field of SLA has shown that during communication breakdowns, such as non-understanding of specific words, participants redirect their attention to the problematic words within the interaction. In such cases, participants shift their attention from conversational progress to addressing the meaning of specific words or phrases (Long, 1996). This negotiation typically follows a structured sequence: a **trigger**, which denotes the problematic word usage; an **indicator**, signaling the problem; and a **response**, aimed at resolving it, sometimes followed by a **reaction to the response** to conclude the sequence.

Indicators, the second component of this type of sequence, play a critical role in highlighting perceived problematic word usage. While they are often clarification requests, they can also take other forms, ranging from explicit meta-comments that directly ask for definitions to more implicit cues, such as silences or gestures signaling confusion (Pitzl, 2005; Vasseur et al., 1996). The nature and explicitness of the indicator often shape how the subsequent response will address the issue, determining whether the negotiation of meaning successfully resolves the misunderstanding.

Varonis and Gass (1985) conceptualize these breakdowns as vertical divergences in a horizontal conversational flow, with side-sequences resolving the issue before the conversation resumes. While this model was initially developed in SLA research, its applicability extends to natural, symmetrical conversations, as broader conversational repair patterns often align with this structure. Research related to non-understanding (Myrendal, 2025) similarly observes this pattern, where conversations shift vertically to address specific lexical misunderstandings before seamlessly returning to the main topic.

Larsson and Myrendal (2017a, 2017b) formalize semantic coordination by modeling WMNs through dialogue acts and meaning update functions. Their work synthesizes taxonomies of dialogue acts, distinguishing negotiation sequences based on comprehension difficulties versus disagreement over word choice. By linking these dialogue acts to specific types of semantic updates, they provide a framework for understanding how word meanings evolve dynamically in interaction.

Noble et al. (2021) build on these concepts by formalizing and annotating WMN as a structured interaction game with defined rules and semantic update functions. Their study uniquely focuses on WMNs in social media contexts, using a regular expression to identify potential sequences on Twitter, such as those containing "What do you mean by X?". Annotators categorized WMNs as stemming from non-understanding or disagreement, aligning with prior models but emphasizing their utility in online discussions. Noble et al.'s work addresses the gap in prior research by providing a formal framework and annotation schema specifically for WMNs in online interaction, highlighting their role in semantic alignment and change.

Despite the insights from prior studies into conversational repair, clarification requests, and word meaning negotiation, no comprehensive corpus dedicated to WMN has been established until now. Similarly, no existing annotation scheme exists that captures the nuanced distinction between types of WMNs, nor do they address the interplay between situated and potential meanings. Our study fills this gap by introducing the NeWMe corpus (**Ne**gotiating **W**ord **M**eaning), the first large-scale manually annotated corpus specifically designed for the study of WMN in diverse interactional settings. Unlike a traditional corpus that consists of newly collected linguistic data, the NeWMe corpus does not contain original material but instead compiles instances of WMNs and related phenomena from existing corpora. In that way, it functions as a "virtual corpus" that provides standoff annotations and tools for accessing and analyzing WMN instances in a structured and convenient way. This resource lays the groundwork for future quantitative and qualitative research into semantic alignment and the dynamics of meaning negotiation. By building on and extending the theoretical frameworks outlined in this section, our contribution advances the empirical study of WMN, providing a gold standard for analyzing these interactional phenomena.

3 Word meaning negotiation

In this section, we introduce the established concept of Word Meaning Negotiation (WMN), drawing on previous research to provide a theoretical basis for our work. While this section does not present a novel contribution, it is essential for framing the subsequent presentation of our annotation scheme and analysis. WMN, as defined by Myrendal (2015, p 19), refers to "instances in communication where participants explicitly negotiate between themselves their respective takes on the situated meaning of a particular word, and/or the meaning potential of that word."

A core component of WMN is the *meta-linguistic shift*, formally defined as a conversational transition in which participants move from discussing the primary topic of interaction to explicitly addressing and negotiating the meaning, interpretation, or use of a specific word or phrase within the interaction. This shift aligns with Clark's (1996) notion of communication occurring on two parallel tracks. Track 1 deals with the "official business" of the conversation, focusing on the primary topic under discussion. Track 2, on the other hand, handles the meta-communicative aspects of interaction, ensuring the conversation runs smoothly by addressing coordination, turn-taking, and grounding. WMN operates within this second track, as it revolves around managing lexical and semantic issues that arise in the course of conversation. By engaging in WMN, participants collaboratively address and negotiate ambiguities or disagreements about word meanings, enabling the conversation on Track 1 to proceed effectively. This interplay between the two tracks highlights the dynamic nature of communication, where both the content and the process are managed simultaneously.

The minimal WMN sequence contains a three-turn exchange, Trigger-Indicator-Response (T-I-R), where the first and second turns are produced by two different people.

1. **Trigger Turn**: This is the initial use of the word that later becomes the focus of negotiation. The word might be a single word form or an expression (e.g., "kick the bucket"). The lexical items that trigger WMNs in our corpus vary widely, ranging from uncommon or technical terms to vague evaluative adjectives and ideologically contested concepts. A more detailed discussion of observed trigger characteristics is presented in Sect. 5.1.
2. **Indicator Turn**: Following the trigger, this turn signals the need to discuss or clarify the word's meaning. It may come in the form of a direct request for clarification of meaning (a meta-linguistic clarification request) or as a challenge to the appropriateness or meaning of the word in the given context (a meta-linguistic objection).
3. **Response Turn(s)**: This turn reflects a meta-linguistic shift, where the discussion moves from the initial topic to focus explicitly on the meaning of the word in question. The shift may not be completely separate from the original discussion topic, but clearly highlights the word's meaning as a central element of the exchange.

WMN sequences may not always occur in immediate chronological succession. In asynchronous online interaction, the turns of a T-I-R sequence can be separated by unrelated posts or delays in participation, yet still be interactionally connected. Recent work on WMN in online forums (Myrendal, 2025) demonstrates that participants use quoting, direct replies, and explicit references to preserve sequential coherence even when turns are chronologically separated. This means that both Indicators and Responses can appear several turns after the Trigger, and the WMN sequence remains intact as long as the interactional meta-linguistic focus on word meaning is maintained. This is also possible in oral interaction if, for example, a speaker detects a misalignment a few turns after the Trigger.

Importantly, WMNs do not always need to result in agreement about the word's meaning to be considered WMNs. Even if participants fail to fully understand or align on a shared interpretation of the meaning of the trigger word, the sequence still counts as a WMN as long as the conversational focus temporarily centers on negotiating the meaning of the word.

WMNs can be initiated in different ways, but often originate in either insufficient understanding of word meaning (this type of WMN is called a NON, short for "non-understanding WMN") or disagreement about what a word can or should mean in a specific situation (this type of WMN is called a DIN, short for "disagreement WMN") (Myrendal, 2015).

Example of WMN Caused by Non-Understanding of Word Meaning (NON)

S1: I'm going to the doctor to get a full body scan tomorrow.
S2: What do you mean by full body scan?
S1: I mean a kind of X-ray where they can see all of the inflamed parts.

This example, taken from Myrendal (2015), illustrates a scenario where S1's use of the term "full body scan" serves as the **trigger**, introducing a word which is not

fully understandable to S2. S2 then produces an **indicator**, explicitly requesting clarification about the meaning of "full body scan," making this phrase the trigger. In **response**, S1 provides an explanation, elaborating on the word to address the lack of understanding. This sequence demonstrates how WMNs initiated by non-understanding (NONs) focus on clarifying the meaning of specific terms to maintain mutual understanding in the conversation.

Example of WMN Caused by Disagreement about Word Meaning (DIN)

> **S1:** Telling children about Santa Claus is straight up lying to them.
> **S2:** That's not what lying means at all!
> **S1:** Of course it is, lying means not telling the truth and everyone knows Santa doesn't exist.

This example, drawn from Myrendal and Larsson (2025), illustrates a WMN caused by disagreement about word meaning (DIN), where the focus shifts to negotiating differing perspectives on the meaning of a word. Here, S1's initial statement introduces the word "lying," which serves as the **trigger**. S2 challenges this usage by providing an **indicator**, asserting that the term "lying" does not apply in the given context and objecting to its use. In **response**, S1 elaborates on their understanding of the word, reinforcing their interpretation and connecting it to the situation at hand.

Once a WMN has been started, the focus of the negotiation may center on either the semantic properties of the word itself—what Norén and Linell (2007) refer to as "meaning potentials"—or the specific details of the situated context in which the word is used. Meaning potentials, as defined by Norén and Linell (2007), encompass the set of abstract properties of a word that, when combined with contextual factors such as linguistic co-text and situational conditions, make possible all reasonably correct or contextually plausible interpretations of that word. These meaning potentials are not static dictionary definitions but dynamic properties shaped by the accumulated interactional experiences of both individuals and the broader language community over time. In contrast, situated meanings emerge in real-time interaction, where the meaning potential of a word is activated and shaped by the immediate context. While meaning potentials provide the flexible foundation of a word's possible uses, situated meanings are collaboratively constructed and negotiated in situ by conversational participants.

WMNs can focus on these two main aspects of meaning: **potential meaning** and **situated meaning**. When the focus is on potential meaning, participants discuss what a word can mean in a broader sense, beyond the immediate context. This might occur in NONs if S2 does not know the word at all or is unfamiliar with the sense in which it is being used (even if it is a well-established sense). In DINs, potential meaning becomes the focus if S2 challenges S1's usage, arguing that it does not align with any established meaning of the word.

On the other hand, WMNs may center on situated meaning, where the negotiation revolves around what the word means in the specific situation being discussed, with less emphasis on its broader usage. In NONs, this could involve S2 struggling to understand how the word applies to the particular object or scenario at hand. In DINs,

situated meaning is at the core when S2 points out that S1's interpretation does not fit the specifics of the current situation.

WMNs can also simultaneously center on **both** potential meanings and situated meanings. In such cases, the negotiation explicitly addresses both what a word can mean in a broader, general sense and what it means in the specific context of the conversation. For instance, S2 might confront S1 by referring to a general definition of the word (e.g., from a dictionary), emphasizing its potential meaning, while S1 explains their intended meaning within the specific situation, focusing on its situated use. Alternatively, S2 might argue that S1's use of the word is unconventional in the given context (situated meaning), prompting S1 to defend their usage by appealing to its broader, standard interpretation (potential meaning). While situated aspects of meaning are always inherently present—since conversations always occur within specific contexts—we consider that WMNs explicitly focus on both dimensions when the negotiation clearly addresses both what the word means in the immediate situation and what it can mean beyond that context.

To clarify how this distinction plays out in practice, we provide three illustrative examples below, one for each type of meaning orientation. These are fictional but representative cases modeled on our corpus data. For further information, we refer the reader to the annotation guidelines (linked in Sect. 7).

Potential meaning

S1: I never considered myself nihilist.
S2: I don't really understand what that word means…
S1: A nihilist is someone who believes that life has no inherent meaning or value.

This is a NON case in which S1's response orients to the potential meaning of the word "nihilist," offering a general, context-independent definition rather than clarifying how it applies in the situation.

Situated meaning

S1: I just think the second method we tried was less effective than the first one.
S2: What do you mean by effective in this case?
S1: I mean it took longer to carry out and didn't give us clear results like the first one did. The first method gave us what we needed almost immediately, but with the second one we had to do extra steps and still weren't sure about the outcome.

This is a NON case in which S1's response orients to the situated meaning of effective. Rather than offering a general definition, S1 clarifies what they meant by the term specifically in relation to the methods being compared in the immediate context.

Both potential and situated meaning

S1: I just think he's being a hypocrite for criticizing people who take government aid.

S2: Isn't a hypocrite someone who pretends to have moral standards but doesn't actually follow them?
S1: Yes, that's what I mean—a hypocrite is someone who publicly upholds a certain standard but then acts against it. And in this case, he's constantly posting about how people should be self-reliant and not depend on welfare, but he accepted unemployment benefits last year without mentioning it.

This is a DIN case in which S1's response clearly draws on both the potential and situated meaning of hypocrite. S1 first confirms the general, context-independent definition (potential meaning), then applies it to the specific situation under discussion (situated meaning).

4 Corpus construction

In this section, we outline the methodology used to create the corpus. Following the selection of a diverse sample of corpora for annotation, we employed a regular expression-based approach to pre-select utterances likely to contain WMN indicators (Sect. 4.1). Section 4.2 introduces the annotation labels and provides illustrative examples for each, highlighting their application in capturing different types of WMNs and related phenomena. We also outline the annotation interface and process, which was conducted by two authors of this paper, detailing the creation of comprehensive annotation guidelines and the calculation of inter-annotator agreement (IAA) to ensure reliability.

4.1 Data selection

4.1.1 Corpora used

The NeWMe corpus consists of a collection of annotated WMNs and related phenomena from existing corpora. In selecting the corpora to be included in the NeWMe corpus, we prioritized sources that offer a range of interactional settings and conversational structures, enabling analyses of WMN across varied contexts. Our aim was to include both spoken and written interactions, spanning dyadic spoken conversations, multiparty spoken interactions, and written online multiparty discussions. This variety ensures a broad representation of how WMNs manifest across different interactional settings, including debates, informal conversations, and other diverse topics and situations. We selected three corpora for annotation:

- The **Switchboard Dialog Act Corpus** (Stolcke et al., 2000) contains 1 155 dyadic phone conversations. Dialogue participants, all from the United States, did not know each other, and were assigned a topic to discuss (e.g., movies, food, vacation spots, recycling, etc.). We access the corpus through the ConvoKit Python library (Chang et al., 2020).
- The **British National Corpus (BNC)** (BNC Consortium, 2007). We use a portion of the spoken part of the BNC, corresponding to 730 spoken interactions.

The data consists of transcripts of conversations recorded in different conversational settings, including lectures, meetings, interviews, medical consultations and news broadcasts, among others.

- The **Winning Arguments (ChangeMyView) Corpus** (Tan et al., 2016), hereafter referred to as **Reddit**, contains conversations from the r/ChangeMyView subreddit posted between 2013 and 2015. In this subreddit, the original poster of a thread states their view on some topic and challenges other users to try to change their opinion, which often generates debate-like conversations with users exposing their arguments. We chose this corpus because we expect lexico-semantic alignment to be particularly important in such a setting: users need to make sure that they are debating about the same things, and disagreements are likely to surface. Reddit contains 3 051 conversations and is also available as part of the ConvoKit package. To facilitate annotation, we consider 586 conversations in which no post or user has been deleted. A specificity of Reddit is its hierarchical structure, with parallel subthreads. We treat each post as an utterance.

4.1.2 Candidate selection with regular expressions

As discussed in Sect. 2, previous research used regular expression matching to identify potential WMNs in Twitter data (Noble et al., 2021). This approach focused on the indicator turn, which typically exhibits less variation compared to the trigger and negotiation phases of a WMN. However, the regular expression employed by Noble et al. was limited to simple variations of the question "What do you mean by X?", such as "what do you actually mean by." In this study, we developed a more extensive list of regular expressions, hypothesizing that they could better capture the diversity of indicator turns.

As a first step towards choosing relevant expressions, we took advantage of the dialogue act annotations in the Switchboard corpus. We compiled all utterances which were tagged with a dialogue act label involving a question (excluding rhetorical questions) or a signal of non-understanding.[2] Through manual inspection of the first 3 000 utterances tagged with these labels, we identified various expressions such as "what is," "what is the difference between," and repetitions of words and phrases from the immediately preceding turn, followed by a question mark (e.g., S1: "I do aerobics, step classes, and…" S2: "step classes?"), which we refer to as the "repetition pattern."[3] We supplemented this list with expressions based on our own intuitions (e.g., "can you define") and insights from prior research (e.g., "this is not X") (Myrendal, 2015).

We used a total of 30 regular expressions alongside the repetition pattern.[4] The complete list of simplified expressions is presented in Sect. 6.2. It includes variations

[2] We used the tags *qy*, *qw*, *qy^d*, *qo*, *br*, *qw^d* and *^g* (http://compprag.christopherpotts.net/swda.html) (accessed January 29, 2025).

[3] For this pattern, we excluded repetitions consisting of "yes ?", "yeah ?" and "no ?", as we observed they had a very large number of matches, especially in the BNC, but were never indicative of a WMN.

[4] The full set of regular expressions and associated retrieval scripts is available in the project's GitHub repository: https://github.com/gu-wmn/NeWMe/.

of expressions such as “what X means”, “in what way/sense?”, “the meaning of”, and others.

Certain adaptations were necessary to take into account the formatting specificities of the included corpora and ensure that the regular expressions would match all relevant candidates. Specifically, several regular expressions were adjusted for Switchboard by making question marks optional, since we observed that they are sometimes omitted (see Example 1 in Sect. 4.2.1). Additionally, before checking for matches on Switchboard, we simplified utterances by temporarily removing indications of laughter, information in curly brackets, and transcription conventions related to spoken phenomena, such as symbols like “+" or “–”. For simplicity, we chose to normalize the transcriptions rather than adjusting the regular expressions, which is more error-prone. For the Reddit data, the regular expression search excluded all cited passages (i.e., where a user quotes text from a previous message), as any relevant matches were already captured in the original posts.

We examined every utterance[5] of the three aforementioned corpora for regular expression matches. Any utterance that matched at least one regular expression was retained as an annotation instance. This procedure obtained 1 305 candidate instances from Switchboard, 2 976 from Reddit and 4 032 from the BNC.

4.2 Annotation

4.2.1 Labels and spans

This section presents the final set of annotation labels used in this study. The annotation process for each candidate began with identifying the type of phenomenon present, assigning a corresponding phenomenon label. Based on this classification, relevant portions of text were marked with span labels, each representing a distinct component of a WMN. When an instance was annotated as a WMN, we additionally annotated the kind of meaning that was being discussed. More information and details can be found in the annotation guidelines, which are provided alongside the annotations.

Types of complete WMNs: NON, DIN and Other

As introduced in Sect. 3, we distinguish between two main types of WMN: **NONs** (non-understanding WMNs), driven by insufficient understanding that prompts S2 to request clarification; and **DINs** (disagreement WMNs), where S2 challenges or objects to S1's use of a word, focusing the negotiation on resolving the disagreement. During annotation, we also identified a third category, **WMN: Other**, for cases where word meaning was discussed without non-understanding or disagreement. These typically involve situations where S2 asks about a word and/or suggests an alternative word, which S1 then confirms as appropriate.

[5] We excluded the first utterance (the title, in Reddit), as the regular expressions should help find potential indicators, which, in a WMN, must happen in a turn posterior to the trigger turn.

Examples of each type of WMN are shown in Examples 1, 2 and 3.[6]

S1: But the problem is, that I am very ***liberal*** politically and so I hardly ever have anybody that wins that I vote for
S2: **Oh, liberal, by, what do you mean by liberal, um**
S1: *Liberal politically, I'm, you know, like pretty left wing Democrat, so*
S2: *Well see, I don't know anything about politics.*
S1: *Oh you don't ?*
S2: *Uh, what's the main difference between Republicans and Democrats?*
(...)
S1: (...) *Democrats usually are more supportive of public assistance programs (...) the big Republican thing is that they don't, they vote for less government, they want less government involvement in society* (...)

Example 1 NON from Switchboard about the word "liberal".

S1: True ***waffles*** are crisp on the outside and fluffy on the inside, and at least double the height of a pancake.
S2: *Let us look at the Merriam-Webster waffle definition. lwaf·fle \□wä-fəl, □w□-\ noun :a crisp cake of batter baked in a waffle iron* (...) **Well which definition are we supposed to go by according to your view? The dictionary definition mentions nothing about waffles being fluffy on the inside or being at least double the height of a pancake, yet your definition mentions bother of these things. So, which definition of a waffle should we be going by?**
S1: *True, the pancake height was my own personal addition for dramatic effect, but that doesn't take away from the fact that Waffle House waffles are not crisp, and therefore do not fit the definition.*
(...)
S2: *So should I go by the dictionary definition of a waffle or your personal definition? If crispness is the sole determining factor then it seems like Waffle House may simply undercook their waffles, not that they aren't waffles at all.*

Example 2 DIN from Reddit about the word "waffle".

S1: Mhm . Do you so you see yourself the fact that you 're living here on your own, in the in the flats , makes you feel more ***vulnerable*** . [UNCLEAR] feel more ***vulnerable*** [UNCLEAR] living in here on your own ?
S2: **You mean frightened ?**
S1: *Yeah I think yeah.*

Example 3 WMN: OTHER from the BNC about the word "vulnerable".

For WMNs, we also annotated the type of meaning. As described in Sect. 3, this could be potential meaning, situated meaning, or both. Example 1 above focuses on potential meaning, Example 2 discusses both kinds of meaning, and Example 3 addresses situated meaning.

Incomplete WMNs and Distractors

While WMNs involve specific sequences where participants explicitly negotiate word meanings, related phenomena may only partially meet these criteria or exhibit distinct but similar interactional patterns. During annotation, we identified such cases

[6] To improve readability, excerpts are formatted so that triggers are marked in bold italics with a green highlight, indicators in bold with a red highlight, and negotiations in italics with a blue highlight. We use "(…)" to indicate omitted portions of the conversation, including partial or full turns, that are not relevant to the example for the sake of brevity.

as either *incomplete WMNs* (Self-Initiated Meaning Negotiation, Non-pursued, Without Trigger) or *distractors* (Reference/NE, Other kinds of clarification requests), serving as negative examples for distinguishing true WMNs. Below, we outline these categories and their defining characteristics and show examples from the corpus:

Self-Initiated Meaning Negotiation (SIMN) occurs when the same person produces both the trigger word and the indicator. For example, the speaker may ask someone else to clarify the meaning of a word just used, to see if that word was understood.

S1: Not a good idea. Erm have you heard of ***half life*** ?
S2: Yeah .
S1: **What does that mean ?**
S2: *that, that's how long it takes for half of the radioactive isotopes to disappear.*
S1: *Great, yeah, that 's a wonderful definition .*

Example 4 A SIMN from the BNC about "half life".

Non-pursued sequences occur when an attempt to initiate a WMN is not successful, i.e., when there is no uptake on the indicator. These cases involve a trigger and an indicator but lack a response turn addressing the semantic issue (i.e., there is no meta-linguistic shift). There is an attempt to initiate a negotiation targeting word meaning, but it does not materialize.

S1: (...) I am telling you race is not a ***scientific*** description and its use as a popular culture description causes hate and ignorance. (...)
S2:(...) **Do you know what you mean by "unscientific", or is it just a non-declarative placeholder speech act you use when you have attitudes of aversion?** (...)
S1: Wow you are putting alot of effort into avoiding my question. I was careful to put it into the most basic terms and yet there was still a communication failure, no problem, lets break out the crayola's. Do you hold the position that race is a scientific descriptor? Your answer should be framed as a "yes", "no" or "unknown". Seriously, dont over think this one, just a one word answer. You know what, let me pull it out of this paragraph and restate, again, no tricks, just a simple one word answer is all that required. Also, if you are worried about tricks go ahead and ask any questions you need, look up any references for clarification, just dont let it interfere with your response.\n\nDo you hold the position that race is an accepted scientific descriptor? Your answer should be framed as a "yes", "no" or "unknown".

Example 5 A Non-pursued sequence from Reddit about the word "(un)scientific." In this case, the utterance containing the indicator receives an answer by S1, but the answer does not address the indicator.

In sequences categorized as **WITHOUT TRIGGER**, the problematic word or phrase is introduced directly in the indicator sentence, without any prior use of the word in the conversation. As a result, the initial trigger component of the three-part exchange required for a sequence to qualify as a WMN (T-I-R) is missing. This type of negotiation may occur when the need to clarify a word's meaning emerges from external factors or prior knowledge (e.g., an external resource, a speaker's experience, or the situational context) rather than from the immediate dialogue. It can also arise when the trigger word is not explicitly used but the conversation addresses a related topic (e.g., S1 discusses feminism without mentioning the word, prompting S2 to indicate a different understanding of "feminism").

S1: (...) what do you think about this new menu for the canteen at Digby's Ballbearings ?
S2: Crunchy nut salad, t , **what 's *tortellini* ?**
S1: *Pasta, stuffed with spinach and cheese* (..)

Example 6 A **WITHOUT TRIGGER** sequence from the BNC.

Distractors involve sequences that resemble WMNs but do not meet the criteria because they do not contain negotiation of word meaning. **REFERENCE/NAMED ENTITY (NE)** sequences focus on resolving referential ambiguity rather than negotiating the meaning of a word (Example 7). Sequences identified as **OTHER KINDS OF CLARIFICATION** requests involve clarification of broader semantic content rather than a specific word's meaning. Often, the request targets the entire utterance rather than an individual lexical item (Example 8).

S1: they'd have their smog alerts where you'd have to stay indoors for so many hours with an air conditioner. And, of course, they don't have ***that*** here in Texas. So, there's…
S2: **You mean they don't have the, uh, the smog alerts?**

Example 7 A REFERENCE/NAMED ENTITY sequence from Switchboard

S1: For this to change my view I'd need strong evidence that presumably lower interest rates on lons from banks have lead to a substantial increase in quality of life as they have increased.
S2: **I'm not sure what you mean here.** Lower interest rates on a loan mean that your monthly payments are lower, and it's easier to pay off / you have more money available for other things. That's a quality of life increase.

Example 8 An OTHER KIND OF CLARIFICATION REQUEST from Reddit

Other labels:NOTHING and IMPOSSIBLE TO ANNOTATE

When none of the above labels applied, an instance was labeled as **NOTHING**. If the annotator suspected that a phenomenon was present but could not be sure due to transcription errors or incompleteness, the instance was labeled as **IMPOSSIBLE TO ANNOTATE**: In the example below from the BNC, the turn that may have contained the trigger could not be fully transcribed:

S1: [UNCLEAR] was n't [UNCLEAR]
S2: Yes, that 's quite nice. So it 's dreamlike what do you mean by that. What 's he doing then ?
S1: Well it 's not very clear whether he 's actually thinking or seeing it.

Example 9 An instance from the BNC labeled as **IMPOSSIBLE TO ANNOTATE.**

Spans: trigger, indicator and negotiation

In our annotation process, a span refers to a specific segment of an utterance that is chosen and labeled with one of three categories: Trigger, Indicator, or Negotiation. When the annotator determined that there was a (complete or incomplete) WMN sequence

or a distractor, the spans corresponding to the three categories were annotated. The **trigger** span was used to tag all problematic word usages, and was annotated for all phenomena except **OTHER KINDS OF CLARIFICATION REQUESTS,** where the request does not target a word or phrase. The **indicator** span was used to highlight the utterance(s) indicating the problem and was used for all phenomena. Finally, the **negotiation** span was used to demarcate the response turns addressing a problem with word meaning, if present; as well as extensive indicator turns containing metalinguistic remarks (see Example 2).

4.2.2 Annotation interface

We used the LabelStudio[7] platform to perform annotation. An instance of annotation consists of a full conversation where one utterance has been matched by one or more regular expressions. The strings matched by the regular expressions were highlighted as potential indicators. The annotator needed to determine whether an instance contained a WMN or some related phenomenon, and, if necessary, select the spans corresponding to its different parts. The annotation focused around the utterance containing the potential indicator or in its close vicinity.[8] A screen capture of the annotation interface is presented in the Appendix (Fig. 1).

4.2.3 Annotation procedure

The annotation was carried out by two expert annotators (the authors of this paper). One author annotated around 81% of the BNC candidate instances and another author annotated the Reddit and Switchboard data as well as the remaining portion of the BNC data.[9] The annotators worked individually but held regular meetings to discuss progress as well as complicated or unexpected cases and ensure consistent application of the schema across the full dataset. These meetings helped not only to ensure consistency in annotation, but also to refine and adapt the set of labels based on the observed data. The decisions taken during these meetings were documented in the annotation guidelines, which were written and updated concurrently.

Once the first iteration of annotation was complete, there was a second round where the two authors went through all instances that had been labeled with some phenomenon and ensured that they were consistent with the final set of labels and their definitions as described in the guidelines. At the end of this process, a stratified sample of 254 candidate instances (175 and 79 by each annotator) was randomly compiled, covering all three annotation categories (WMN, Distractor, Nothing) and including data from all three corpora (BNC, Switchboard, and Reddit). These instances were manually reviewed by the other author to assess agreement. In cases

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

[8] We also considered phenomena taking place in other parts of the conversation if they were detected by the annotators while reading the context to understand the situation.

[9] The distribution of annotations was driven by practical considerations and differences in corpus format. This decision favored intra-corpus consistency, while inter-corpus consistency is supported through the regular meetings and the agreement study described in this section.

of disagreement, the two authors discussed the annotations in an effort to reach consensus. However, in a small number of instances, consensus could not be achieved, and the original annotator's choice was kept. Most disagreements stemmed from differing interpretations of the specific conversational context. WMN annotation is an inherently subjective task and the presence of some disagreement is expected.[10] Some form of disagreement remained only in 14 instances of the sample, 10 of which concerned the label used, and 4 others about the type of meaning, resulting in an agreement of 94–96% between the two authors, compared to a 86–90% agreement before discussion. Examples 10 and 11 illustrate two cases of disagreement between annotators.

S1: A copied ***product*** violates the right of the creator to use their product in any way they want so as to reach their clients.
S2: **Not if "the product" means the idea in his mind and not the copies.** Interfering with the creator's notes or thoughts, for example, would be unjust. But copying his work creates something that is not "their product". In general English, "product" can mean "the result of" or "something made or grown to be sold". One meaning confers ownership, the other doesn't.
S1: *It seems that we have a disagreement about whether the ideas are the same thing or not* (...) *But we are talking about taking the idea and creating essentially the same product, meaning that you have interpreted it in exactly the same way.*(...)

Example 10 A Reddit instance where expert annotators disagreed on the type of label (WMN: OTHER or DIN). One author viewed this as proposing an alternative interpretation of the word "product", the other saw it as a disagreement on its meaning. The original annotation (WMN: OTHER) was kept.

S1: (...) but there are categories where the diamond is superior. Carbon content by mass, hardness, and ***price*** come to mind. (...)
S2: (...) 3. **What do you mean by price - cheaper or less expensive?** *Because diamonds can be beat easily by a multitude of other stones in terms of total cost, and they're certainly not cheap. One meaning confers ownership, the other doesn't.*
S1: I'll concede this point, I'm not an expert by far at buying or selling gemstones and was unaware of their position on the gradient of prices.

Example 11 A Reddit instance where expert annotators were in disagreement. While both annotators agreed that S2 confronts the use of the word "price", one author saw this as a question on topic rather than an attempt to learn more about the meaning of the word in this context. This was finally annotated as a NON-PURSUED.

The speed of annotation varied depending on the corpus, how many interesting cases were found, and the acquired experience both with the annotation task and the LabelStudio platform. Uninteresting cases could often be quickly detected and annotated as NOTHING. Reddit data had the additional complexity of having to keep track of the response subthreads. The average annotation speed was estimated to be 1.4 candidate instances per minute for one author and 1.05 for the other.

[10] While recent work in NLP is increasingly embracing *perspectivist* approaches to annotation (Frenda et al., 2024), which encourage subjectivity, we adopt here a *prescriptive* protocol (Rötger et al., 2022), centered on the consistent application of annotation guidelines. This supports our goal of creating a resource that enables further study of WMN by delimiting this phenomenon as clearly as possible.

Table 1 Frequency of every phenomenon in the WMN corpus by source

	BNC	Reddit	Switchboard	Total
NONs	116	66	33	215
DINs	11	158	0	169
WMN: Other	14	3	3	20
Non-pursued	4	197	2	203
SIMN	37	2	3	42
Without trigger	10	0	2	12
Reference/NE	7	3	18	28
Other kinds of clarification requests	15	109	49	173
Nothing	3746	2353	1188	7287
Impossible to annotate	24	1	0	25
Total	3984	2892	1298	8174
Total phenomena	214	538	110	721

5 Analysis

5.1 Corpus statistics

Table 1 presents the number of phenomena found in the WMN corpus. We annotated a total of 404 WMNs: 215 NONs, 169 DINs and 20 WMN: Other. WMN types are not evenly distributed across sources: A large majority of DINs (94%) are found in Reddit data. This aligns with findings that anonymous online platforms often foster interest-based selectivity and reduced multidimensionality in social affiliations, which can amplify disagreement (Barnidge, 2017). Unlike social media networks that articulate connections across diverse social dimensions (boyd & Ellison, 2007; Ellison et al., 2007), anonymous forums lack shared common ground and non-verbal cues, making disagreement more frequent and fostering environments conducive to word meaning negotiation. In contrast, most NONs (69%) are found in spoken conversation. This could also be expected, as the asynchronicity of online forums gives participants time to reflect upon or look up word meanings (Myrendal, 2019); while the immediacy of spoken communication requires solving the understanding problems on the spot in order to successfully continue communication. With only 20 instances in the whole corpus, WMN: Other was not very common, but not an isolated case either. The lower incidence of WMNs in Switchboard can be explained by its smaller size: Although it contains more conversations than the BNC and Reddit portions considered, conversations and utterances are typically much shorter. This is reflected in the token counts: while the BNC and Reddit portions contain approximately 6 and 3 million tokens, Switchboard contains fewer than 2 million.

The distribution of incomplete WMNs also paints an interesting picture. NON-PURSUED sequences are almost exclusively (97%) found in Reddit data, and they are more numerous than DINs in this subset of the data (197 vs 158). This makes sense, as it is much easier to stop a thread, ignore a post or change the topic in online forums than it is to ignore a clarification request in real-time spoken conversation. In spoken interaction, research within Conversation Analysis (CA) has established a strong preference for an answer to come in the next turn when a direct question is asked (Heritage, 1984;

Sacks et al., 1974). Questions are typically treated as making a response conditionally relevant, and failing to provide an answer is often interpreted as a breach of normative expectations. This orientation underscores the importance of maintaining interactional continuity and progressivity in spoken interaction (Schegloff, 2007; Stivers & Robinson, 2006). Furthermore, Karafoti (2021) shows that delayed or missing responses stand out and often require extra effort, like rephrasing or explaining, to repair. In contrast, asynchronous online forums like Reddit lack the real-time pressure of spoken interaction, allowing participants to ignore posts or disengage from threads without facing social consequences. This flexibility leads to a higher number of NON-PURSUED WMNs in Reddit data, where many posts (70%) remain unanswered or receive responses that don't engage in word meaning negotiation (30%). Unlike face-to-face communication, where conversations rely on immediate and accountable responses (Karafoti, 2021; Pomerantz & Heritage, 2013; Schegloff, 2007), online interaction reduces the obligation to reply, making non-responses less noticeable and less socially significant.

In contrast to the NON-PURSUED sequences, SELF-INITIATED MEANING NEGOTIATION (SIMNs) were most commonly found in the BNC data. These cases often involve classroom-like interactions, where a teacher introduces a word and asks students if they know its meaning, prompting a meta-linguistic discussion for educational purposes. Instances Without trigger were not very common (12 in total). These instances typically arise when the problematic word comes from a source external to the conversation (e.g., someone is reading something and finds a word they don't understand), as well as in teacher-student interactions. With 173 instances, OTHER KINDS OF CLARIFICATION REQUESTS were among the most commonly encountered phenomena, especially in the Reddit data. In contrast, the other distractor, REFERENCE/NE was less frequent, appearing only 28 times.

We also want to draw attention to the large number of irrelevant cases, annotated as NOTHING, which provides insight into the effectiveness of the regular expressions used to capture relevant phenomena.[11] Only about 5% and 8% (respectively) of the retrieved BNC and Switchboard candidate instances contained interesting phenomena, compared to 19% of Reddit. However, this drops to 8% for Reddit when considering only complete WMN sequences. While NOTHING instances are often quick to annotate, these percentages point to a clear opportunity for improvement. Future work could focus on refining the regular expressions or using semi-automatic methods to optimize and speed up the annotation process, pre-selecting candidate instances that are more likely to contain a relevant phenomenon (see Sect. 5.2 for more details).

To provide a clearer picture of the prevalence of identified WMNs in our data, we note that regular expressions returned matches in 1863 distinct conversations (75% of the 2 471 conversations considered), but only 276 of these contained complete WMNs (11% of all available conversations). Looking at these results by corpus, Reddit was the most dense in WMN: 25% of all Reddit conversations contained at least one WMN, compared to 13% and 3% in BNC and Switchboard, respectively. The lower incidence rate in Switchboard may be due to the nature of the interactions. The debate-like nature of Reddit is particularly conducive to discussing word meaning, while Switchboard conversations are casual, short, and do not have a clear communicative goal other than

[11] We exclude from these corpus statistics any instances where a phenomenon was present but had already been annotated in another instance from the same conversation.

Table 2 Frequency of the type of meaning by WMN type and corpus source

	BNC	Reddit	Switchboard	Total
NON				
Situated	88 (75.9%)	52 (78.8%)	17 (51.5%)	157 (73.0%)
Potential	20 (17.2%)	8 (12.1%)	13 (39.4%)	41 (19.1%)
Both	8 (6.9%)	6 (9.1%)	3 (9.1%)	17 (7.9%)
DIN				
Situated	5 (45.5%)	32 (20.3%)	0	37 (21.9%)
Potential	0	58 (36.7%)	0	58 (34.3%)
Both	6 (54.5%)	68 (43.0%)	0	74 (43.8%)
Other				
Situated	14 (100%)	2 (66.7%)	3 (100%)	19 (95.0%)
Potential	0	0	0	0
Both	0	1 (33.0%)	0	1 (5.0%)
All WMNs				
Situated	107 (75.9%)	86 (37.9%)	20 (55.6%)	213 (52.7%)
Potential	20 (14.2%)	66 (29.1%)	13 (36.1%)	99 (24.5%)
Both	14 (9.9%)	75 (33.0%)	3 (8.3%)	92 (22.8%)

discussing a preselected topic. As a result, clarifying word meaning is less crucial to carry on the conversation. It is important to note, however, that our analysis is limited to candidate instances matched by our regular expressions, and we do not have an estimate of how many WMNs may have been missed.

In Table 2, we present statistics on the type of meaning of each WMN, categorized by source. When examining patterns for each type of WMN, we observe that NONs and WMN: Other most frequently involve discussions about the *situated meaning* of the word. This indicates that insufficient understanding, in our corpus, is mostly tied to the specific way a word is used or its application to a specific situation. WMN: Other interactions often involve corrections or suggestions of alternative lexical choices or phrasing, and thus their tendency to relate to situated meaning is not surprising. DINs, however, show a different pattern: these interactions more often address the *potential meaning* of a word, or a combination of both its situated and potential meanings. In many cases, disagreement arises when participants bring up definitions of words as a way to defend their viewpoints, highlighting the potential meaning of the word. See for instance Example 12 below. Sometimes, post authors use definitions to contrast them with the situated meaning proposed by another participant, which they may perceive as incorrect or inadequate.

S1: (...) you're saying that ***patriotism*** and similar impulses necessarily lead to hatred and discrimination. I think this is untrue, and that there are many people that are both patriotic and deeply empathetic toward people from different countries (...)
S2: **Well what is patriotism then ?!**
S1: *Love of country. Love of one thing doesn't imply or lead to hatred of another. Nor does it require that you be vocal about it.*
S2: *Ok. I'm from eastern europe, so my view is clearly biased here, because the concept of "love of country" seems unnatural to me.*
S1: *Fair enough, but that's the [dictionary definition](http://www.merriam-webster.com/dictionary/patriotism) of patriotism. Thinking it's weird is different than thinking it leads to negative feelings.*

Example 12 A DIN from Reddit discussing the potential meaning of the word "patriotism".

We also examine the characteristics of the annotated spans. For triggers, the majority (71%) consist of individual words, while 27% are multi-word phrases, and 2% involve a combination of both (e.g., "feminism" and "feminist movement"). Interestingly, the spoken corpora contain a higher proportion of phrase triggers, with 36% in the BNC and 39% in Switchboard, compared to 21% in Reddit. An interesting direction for future research would be to explore how the nature of the trigger—whether it is a single word or a (lexicalized) phrase—relates to the type of meaning being negotiated. Our data suggests that WMNs focusing on situated meaning are more likely to involve trigger phrases than those focusing on potential meaning. This pattern may arise because phrases inherently provide more contextual information, anchoring their meaning more directly to the specific utterance in which they appear. For example, in "What do you mean by 'light meal'?", the phrase "light meal" invites a discussion about the meaning of "meal" in this particular context (e.g., portion size or calorie content), reinforcing its situatedness. In contrast, individual words, such as "calculate" in "What do you mean by 'calculate'?", are more likely to invite discussion of their broader, more general meaning. Notably, phrases do not typically have potential meanings unless they are lexicalized expressions. Regarding the extent of negotiation, DINs tend to involve longer exchanges than NONs, averaging 7.2 turns compared to 3.5 turns. This difference aligns with expectations as well: NONs typically conclude once meaning is clarified, while in DINs each participant provides their own perspective and engages in debating word meaning. This pattern is consistent with Myrendal's (2015) original study of WMN, which also found that NONs are typically brief and uniform, concluding within three to four turns on average. In contrast, DINs display much greater variability in length, in Myrendal's study ranging from as few as five turns to as many as 85. In our corpus, the longest NON contains 27 turns in total, while the longest DIN spans 268 turns. This highlights the more elaborate and prolonged nature of DINs, where participants engage in extended exchanges to explore and debate different interpretations of word meaning.

In terms of the trigger words observed during annotation, some general patterns emerged. Several triggers are uncommon or specialized terms such as "abominably", "abstract sum", or "sentience", which may not be part of every speaker's vocabulary. Others involve vague or context-sensitive expressions like "acceptable" or "vulnerable", which often prompt clarification or elaboration due to their evaluative or relative nature. In the Reddit data, we frequently encountered ideologically charged or culturally contested terms that gave rise to disagreement-driven WMNs (DINs). Examples include "patriotism", "religion", "eugenics", "marriage", "war on terror", "magical thinking", and "absolute power". These terms tend to provoke negotiation around what the word "should" mean, often involving appeals to definitions or contrasting ideological stances. While a systematic quantitative analysis of trigger word frequency is beyond the scope of this paper, we highlight these examples to illustrate the lexical variety observed in the corpus. Related to this, Noble et al. (2025) examine the lexical and semantic features of trigger expressions from the NeWMe corpus. The paper analyzes triggers by concreteness, sentiment, part of speech, interaction modality and form, distinguishing patterns between disagreement- and non-under-

standing-driven cases and shedding light on how different kinds of expressions are likely to trigger different kinds of negotiations of meaning in dialogue. One interesting observation is that abstract expressions are associated with disagreement about word meaning, while concrete expressions are relatively more associated with negotiations due to misunderstanding.

Finally, we also look into the number of speakers involved in a WMN in multiparty conversations. In the Reddit portion of NeWMe, we observe that 22% of WMNs involve only two participants. In this respect, there is no substantial difference between NONs (21%) and DINs (22%). This means that, in most cases, there is at least one additional speaker joining in the negotiation. In the BNC, instead, 77% of WMNs in non-dyadic interactions (79 out of 103) involved only two speakers. However, these figures should be interpreted with caution, as the BNC sometimes contains errors in speaker identification. A more fine-grained analysis is left for future work.

5.2 Evaluating the success rate of regular expressions

In this section, we evaluate the effectiveness of our regular expression approach in identifying relevant phenomena and analyze variations in indicator expressions across the three corpora for NONs and DINs. By analyzing the success rates and outcomes of each pattern, we aim to understand their contributions to the annotation process and identify areas for improvement in future methodology refinements.

Table 3 displays the regular expressions (regexes) used to search for potential WMNs, alongside their performance metrics. The regexes are sorted from highest to lowest with regards to how many WMNs they produced in the searches. Each row corresponds to a specific regex, providing the following details:

- **Regex Name**: The name given to the regular expressions used to identify potential WMNs in the dataset.
- **Total Matches**: The number of candidate instances in the dataset that matched the regex.
- **# WMNs**: The count of instances identified as WMNs among the matches.
- **% WMNs**: The proportion of matches that were confirmed as WMNs, reflecting the precision of the regex for WMN identification.
- **# Successful Matches**: Successful matches include both WMNs and other related discursive phenomena of interest to this study. Aside from WMNs, this also includes incomplete WMNs and distractors.
- **% Successful Matches**: The proportion of matches considered relevant to the broader analysis, including confirmed WMNs, incomplete WMNs, and distractors. This metric reflects the regex's overall effectiveness in capturing both strict WMNs and related phenomena of interest.

This table allows for a detailed comparison of the performance of each regex, displaying differences in their specificity and success rates for identifying WMNs or other phenomena considered relevant in this study. Regexes with higher percentages of successful matches (WMN or not) demonstrate greater precision, while those with

Table 3 Overall success of every regular expression

Regex name	Total matches	# WMNs	% WMNs	# Successful matches	% Successful matches
you mean	834	172	20.6	385	46.2
can you define	853	149	17.5	224	26.3
definition of	481	121	25.2	192	39.9
what is	1484	65	4.9	122	8.2
what do you mean by	154	63	40.9	114	74.0
this is not X	1929	59	3.1	118	6.1
is that a/the	952	37	3.9	80	8.4
this isn't X	873	36	4.1	61	7.00
[repetition pattern]	1006	30	3.0	34	3.4
what X means	346	28	8.1	47	13.6
the term	157	20	12.7	28	17.8
in what way/sense	73	11	15.1	19	26.0
what does X mean	108	11	10.2	33	30.6
are you talking about	52	7	13.5	20	38.5
what is the difference between	70	7	10.0	13	18.6
Isn't this…?	78	7	9.0	10	12.8
what sort/kind of	132	6	4.5	8	6.1
elaborate	119	6	5.0	18	15.1
the meaning of	51	5	9.8	8	15.7
is this not…?	35	4	11.4	5	14.3
are we talking	38	3	7.9	9	23.7
do you know what X is	43	2	4.7	10	23.3
is this what you mean	17	2	11.8	8	47.1
word X means	28	2	7.1	3	10.7
have you heard of	40	1	2.5	3	7.5
is that what X is	15	1	6.7	2	13.3
does X count as Y	6	1	16.7	1	16.7
how do you mean that	2	0	0.0	2	100.0
that would mean	48	0	0.0	0	0.0
is that what that/this is	1	0	0.0	0	0.0
depends on what you mean	2	0	0.0	0	0.0

high total matches but lower success rates highlight the trade-off between broad coverage and targeted accuracy.[12,13]

When examining the table of regex patterns and their success rates in identifying instances of WMN, several observations stand out. First, the regex pattern "you mean" generates the highest number of successful matches of the tested expressions

[12] The regular expression "is that a/the" was excluded from the BNC corpus to speed up the annotation process. Early in the reference annotation phase, it became evident that while this expression generated a large number of matches, its success rate was very low. The cases listed in Table 3 for this regular expression in the BNC reflect instances where other regular expressions also matched the same utterance.

[13] Phenomena that were matched by more than one regular expression count as successfully identified instances for each of the matching regular expressions. For this reason, the numbers in this table do not align exactly with the global corpus statistics.

(172), indicating its common usage in interactions where participants initiate WMN by seeking clarification or questioning the meaning of words and phrases. In comparison, the regex "what do you mean by"—which subsumes the broader regex "you mean"—produces fewer successful matches (63) but has a higher precision than the broader regex (40.9% successful matches, compared to 20.6% for "you mean"). The pattern "you mean" is more open-ended and can appear in various contexts that may not always involve a direct question about word meaning, which could lead to a higher number of hits but a lower precision in identifying WMNs. In contrast, "What do you mean by", which has a high precision in identifying WMNs, explicitly starts with an interrogative word followed by a direct address "you", making it more likely to be used in contexts where semantic clarification is sought regarding a word's meaning.

Further down the list, we find some regexes that produce a considerable amount of successful WMNs, but that have a very low precision rate. For example, the regex "what is", "this is not X", "this isn't X", "is that a/the" and the repetition pattern produce between 37 and 65 WMNs, but generate a very high amount of NOTHING. For instance, the regex "what is" generates 65 WMNs, which is significant. However, it also results in 1 484 total hits, indicating a substantial number of irrelevant sequences that researchers must sort through to find the interesting cases. This highlights a common challenge with broader patterns: while they can produce a wide range of potential negotiations, their lower precision requires significant validation which can take a lot of time.

For brevity and ease of presentation, subsequent tables will include only the most productive regular expressions. Full versions of the tables can be found in the Appendix.

When looking at the success rate of each regex in relation to the three different corpora in Table 4, some similarities and differences appear. For example, the regex "you mean" displays varying levels of success across the corpora. In the BNC, it produced 424 matches, with 61 confirmed WMNs, resulting in a precision of 14.4%. In Switchboard, the regex generated 91 matches but achieved a higher precision of 20.9%, with 19 confirmed WMNs. In contrast, this regex performed much better in the Reddit corpus, producing 319 matches with 92 confirmed WMNs, achieving a precision of 28.8%. This strong performance in online argumentative settings highlights the regex's relevance in text-based debates, where participants frequently engage in clarification or semantic negotiation, making it an effective tool for WMN detection in such settings.

Similarly, the two regex patterns explicitly referencing definitions—"can you define" and "definition of"—performed significantly better in the online Reddit corpus compared to the spoken corpora in BNC and Switchboard. "can you define" generated 263 matches in the BNC but only 7 WMNs, resulting in a low precision of 2.7%. Even lower numbers were found in Switchboard, where the regex generated 16 total matches but only 1 WMN (6.2% precision). However, in Reddit, it produced 574 matches and 141 confirmed WMNs, achieving a much higher precision of 24.6%. This strong performance highlights its alignment with the debate-oriented nature of online discourse, where explicit requests for definitions are more common than in spoken interaction. The regex "definition of" displays a similar pattern. In the BNC,

Table 4 Regex success by corpus

Regex name	BNC			Switchboard			Reddit		
	# matches	# WMN	% WMN	# matches	# WMN	% WMN	# matches	# WMN	% WMN
you mean	424	61	14.4	91	19	20.9	319	92	28.8
can you define	263	7	2.7	16	1	6.2	574	141	24.6
definition of	95	4	4.2	2	0	0.0	384	117	30.5
what is	812	35	4.3	439	10	2.3	233	20	8.6
what do you mean by	61	23	37.7	2	2	100.0	91	38	41.8
this is not X	809	8	1.0	244	0	0.0	876	51	5.8
is that a/ the	219	3	1.4	263	1	0.4	470	33	7.0
this isn't X	264	1	0.4	68	2	2.9	541	33	6.1
[repetition pattern]	888	26	2.9	113	4	3.5	5	0	0.0
what X means	202	3	1.5	11	0	0.0	133	25	18.8
the term	90	4	4.4	11	0	0.0	56	16	28.6
in what way/sense	43	3	7.0	5	0	0.0	25	8	32.0
what does X mean	96	9	9.4	0	0	0.0	12	2	16.7
are you talking about	18	0	0.0	16	3	18.8	18	4	22.2
what is the difference between	49	5	10.2	5	2	40.0	16	0	0.0

it generated 95 matches with just 4 WMNs (4.2% precision), indicating limited use in meaningful negotiation within this corpus. Its presence in Switchboard was negligible, with only 2 matches and no confirmed WMNs. However, it performed markedly better in Reddit, where it produced 384 matches and 117 WMNs, achieving a precision of 30.5%. This suggests that "definition of" is particularly suited to capturing instances of semantic clarification and negotiation in text-based online interactions.

Together, these patterns highlight the importance of context in determining regex effectiveness, with explicit mentions of definitions proving far more productive in online settings than in spoken interaction.

The performance of regexes in identifying WMNs varies significantly depending on the type of WMN they capture: NONs (WMNs originating in insufficient/non-understanding of word meaning) and DINs (WMNs arising from disagreement about what a word can or should mean), which is displayed in Table 5. Regexes such as "you mean", "what do you mean by" and "what is" are especially effective at identifying NONs. These patterns often appear in contexts where participants request clarification to address a lack of understanding or to seek further explanation of a

Table 5 Regex success in relation to WMN type

Regex name	# WMNs	# NONs	# DINs	# Other	% NONs	% DINs	% Other
you mean	172	107	45	20	62.2	26.2	11.6
can you define	149	42	107	0	28.2	71.8	0.0
definition of	121	11	110	0	9.1	90.9	0.0
what is	65	46	18	1	70.8	27.7	1.5
what do you mean by	63	46	16	1	73.0	25.4	1.6
this is not X	59	3	55	1	5.1	93.2	1.7
is that a/the	37	9	28	0	24.3	75.7	0.0
this isn't X	36	7	29	0	19.4	80.6	0.0
[repetition pattern]	30	29	0	1	96.7	0.00	3.3
what X means	28	5	23	0	17.9	82.1	0.0
the term	20	4	16	0	20.0	80.0	0.0
in what way/sense	11	3	8	0	27.3%	72.7	0.0
what does X mean	11	10	1	0	90.9	9.1	0.0
are you talking about	7	4	1	2	57.1	14.3	28.6
what is the difference between	7	6	1	0	85.7	14.3	0.0

word's meaning. Their frequent association with NONs suggests that these regexes are well-suited for detecting sequences where comprehension issues are the driving force of the negotiation.

Conversely, regexes such as "can you define" and "definition of" are more strongly associated with DINs, capturing interaction sequences where participants are in disagreement about word meanings and need to draw upon definitions to make their case about what a word actually means. As a result, these patterns tend to appear in discussions that involve challenges to interpretations or efforts to redefine words, making them especially relevant in contexts of semantic negotiation.

In addition, the regexes "this is not X" and "this isn't X" stand out for their strong association with DINs, with almost all of the matches falling into this category. Although these patterns did not produce a high number of total matches, their precision in identifying DINs is exceptionally high, reflecting their specialized role in capturing sequences of negotiation of word meaning originating in disagreement. This connection is not surprising, as these phrases are commonly used as meta-linguistic objections initiating DINs. Such objections occur when a participant challenges the appropriateness or validity of a word used by another participant, asserting that the term does not fit the context or intended meaning. Overall, the strong connection between these regexes and DINs highlights their importance in identifying disagreement-driven negotiations of meaning, making them valuable tools for studying these specific types of interactional phenomena.

The use of regex patterns in identifying WMNs has proven effective but also highlights areas for refinement. Future studies aiming to build robust WMN corpora more efficiently should consider both the strengths and limitations observed in current regex performance. Regexes such as "you mean" and "what do you mean by" illustrate the importance of balancing specificity and coverage. Regexes such as "what is", "this is not X" and the repetition pattern capture valuable WMNs but are overly broad, leading to many irrelevant matches. To address this, future studies could analyze positive and negative cases, i.e., look more closely at both the successfully identified WMNs

and the irrelevant matches produced. Identifying commonalities in the negative cases could perhaps lead to conclusions that could inform the creation of more targeted patterns, leading to less irrelevant matches and more successful candidate instances.

We carried out this analysis for cases matching the repetition pattern and compiled a list of 34 expressions such as "Is it?" or "Really?" which were almost never part of an interesting phenomenon. We recalculated its success rate after excluding all instances with these expressions. The percentage of successfully captured WMNs went from 3 to 3.7%, while that of total successful matches increased from 3.4% to 4.2%. While this may not seem like a substantial increase, and one WMN is lost in the process, 220 instances that were annotated as NOTHING would have been omitted. This is an example of an easy-to-implement improvement that can help speed up the annotation process.

Other improvement methods could involve the addition of contextual constraints, such as multi-turn patterns that take into consideration the turn following a potential indicator. A regex-based approach could also be complemented with machine learning techniques, using the annotated data to train models to classify sequences more accurately.

6 Inter-Annotator Agreement

In this section, we present our inter-annotator agreement (IAA) study. Three annotators were recruited and trained to label a sample of the corpus data. We then evaluated their agreement with one another as well as with our reference annotations.

6.1 Method

6.1.1 Annotation phases

We involved three female Master's students in Computational Linguistics from the University of Gothenburg to annotate a subset of the corpus in order to evaluate the complexity of the annotation task and the clarity of our guidelines. The annotators were native speakers of Polish, Brazilian Portuguese, and Finnish, all with an advanced level of English and fluent speaking skills. This step was crucial to assess how well-defined the task is, and to estimate the reliability that could be expected from the annotation process. The process was divided into three stages: two training phases and the annotation of a sample of the corpus.

Training 1: Preparation and learning

The annotators were introduced to the task through the written guidelines and two training videos. These videos covered the concept of WMN, the use of the Label Studio annotation platform, and included a few illustrative examples. Following this training, we held a joint meeting with all annotators to address any questions or concerns they had on the material before they proceeded to annotate independently.

Training 2: Practice and refinement

In this phase, the annotators worked through two rounds of practice annotations, each consisting of 15 examples. The training samples comprised a total of ten NONs, five DINs, two OTHER KINDS OF CLARIFICATION REQUESTS, one NON-PURSUED sequence and 12 NOTHING. After each round, we provided personalized feedback based on the expected annotations. This process also resulted in enhancements to the guidelines, clarifying areas that had been unclear during the practice annotations.

Main annotation phase

Once training was complete, annotators proceeded to annotate the sample with no further feedback. All three worked on the same data to allow for inter-annotator agreement analysis. No additional instructions were given, except for resolving technical or procedural issues. This hands-off approach ensured that the annotations reflected the annotators' independent understanding of the task.

6.1.2 Sample composition

Table 6 presents the composition of the sample annotated by the three annotators, which consists of 704 candidate instances: 305 from the BNC, 277 from Reddit and 122 from Switchboard. We based our sample selection on the distribution of labels related to the phenomena of interest within our corpus and across the three data sources. To decide the number of phenomena to include, we slightly adjusted proportions according to two criteria: for rare phenomena, such as Without trigger, we included all available cases to ensure the most reliable agreement estimation possible. We also prioritized including complete WMNs over incomplete ones, as they are the focus of our study. After establishing the sample composition, we randomly sampled the number of instances that had been decided upon for each label from the corpus. Finally, we completed the sample by adding the same number of randomly sampled NOTHING sequences as there were instances for all other labels together. The sample size was determined by the anticipated time commitment for the annotators, calculated based on their pace during the training phase. It took the annotators between 57

Table 6 Composition of the sample for student annotators

Expert-annotated label	BNC	Reddit	Switchboard	Total
NONs	53	28	16	97
DINs	10	62	0	72
WMN: Other	14	3	3	20
Non-pursued	3	51	1	55
SIMN	26	2	3	31
Without trigger	10	0	2	12
Reference/NE	6	3	16	25
Other kinds of clarification requests	4	25	11	40
Total relevant phenomena	126	174	52	352
Nothing	179	103	70	352

and 65.5 h to complete the main annotation phase, which corresponds to 0.18 to 0.2 instances per minute. Notably, this sample is much more dense in terms of interesting (not NOTHING) cases compared to the entire NeWMe corpus. Additionally, the annotators received candidate instances in a random sequence, which likely slowed down the annotation process compared to the reference. In the reference setup, NOTHING cases were more common and instances from the same conversation were presented sequentially, offering more context and clarity.

6.1.3 Agreement metrics

Label agreement

We verify the extent to which annotators agree with each other and with the reference regarding the phenomenon (or, in a few cases, phenomena) present in an instance. We calculate this for the raw annotations, as well as at the level of individual labels and sets of labels within specific categories: WMN (NONs, DINs and Other) and WMN-like (which refers to the combination of WMNs and incomplete WMNs: NON-PURSUED, SIMN and WITHOUT TRIGGER). To assess this, we use **Krippendorff's alpha** (Krippendorff, 2013) because of its flexibility in handling the number of annotators, possible values and types of data. This metric ranges from -1 to 1, with 1 indicating perfect agreement between annotators. This agreement is calculated over all available instances. Since some instances were annotated with more than one phenomenon, we modified Krippendorff's alpha difference function to be the inverse of the overlap in a set of labels (i.e., if one annotator found two phenomena in an instance, e.g., a NON-PURSUED and a NON, and the other annotator only found the NON, the difference metric would be 0.5, reflecting the partial agreement between the two annotators).

Span agreement

When annotators agree that a phenomenon is present, we measure their agreement on the location of its components (trigger, indicator and negotiation). Here, we quantify the annotators' agreement with each other and with the reference on each kind of span label. Note that this agreement is only relevant when the spans are present; that is, when the annotators agree that the instance in question contains a phenomenon requiring span annotation. Consequently, for this calculation, we include only instances where both annotators marked the relevant spans.

For **triggers**, we calculate two measures: the **exact span match ratio** and the **string overlap ratio**, which we report averaged across instances. The first measure takes into account the exact location of the triggers annotated, down to the character level. The second measure calculates the overlap in the sets of words annotated as triggers. These two measures are complementary. For example, if two annotators found a phenomenon related to the word "tea," but they annotated the trigger in different utterances, the exact span match ratio will be 0 but the string overlap ratio will be 1, because they agreed on the actual problematic word, but not on its exact location. Conversely, if they agreed on the trigger's location but one of them annotated it as "green tea" instead of "tea," the string overlap ratio would be 0, but the exact span match ratio would be 1/3.

Table 7 Krippendorff's alpha for groups of phenomena (WMN and WMN-like) and for the raw annotation

Annotators	WMN	WMN-like	Raw annotation
A1-R	0.51	0.60	0.56
A2-R	0.63	**0.65**	0.55
A3-R	0.64	**0.68**	0.60
A1-A2-A3	0.53	0.59	0.51
A1-A2-A3-R	0.56	0.62	0.54

For indicators and negotiations, we do not evaluate exact span overlap, but instead focus on agreement at the utterance level. For the **indicator**, we report the percentage of phenomena where both annotators marked the indicator in the same utterance. For the **negotiation**, which can span multiple turns, we calculate the **F1-score** at the utterance level, which measures the extent to which two annotators agree on which utterances of the conversation contained a negotiation span.[14] We report the average F1-score of all instances compared.

Agreement on type of meaning

We calculate **Krippendorff's alpha** to assess the agreement on the type of meaning discussed in a WMN. This calculation is restricted to instances where annotators agreed on the presence of a WMN. We treat the annotation of "both" types of meaning as a double annotation of situated and potential meaning instead of as a separate third label, and we use the inverse overlap metric as done for the label agreement.

6.2 Results

In this section we present all inter-annotator agreement results. We refer to the three annotators as A1, A2 and A3, and to the expert reference annotation as R.

6.2.1 Label agreement

We begin our analysis by examining label agreement across different label groups and on the raw annotation (Table 7) and each individual label (Table 8). The highest global agreement (A1-A2-A3-R) was observed for the NOTHING label, indicating that determining whether there is some interesting phenomenon in an instance or not was the easiest task to agree upon. Interestingly, NONs display a much higher agreement than DINs (0.55 vs 0.35). While this could suggest that DINs are more challenging to annotate consistently, it may also reflect corpus-specific effects, which we describe below.

Agreement varied considerably between annotators. Overall, A3 was the most closely aligned with the reference annotations, achieving up to 0.68 agreement for WMN-like phenomena. Krippendorff's alpha values are often considered to indicate moderate agreement starting at 0.68, reflecting the global difficulty and subjectivity of the task. However, A3's performance, approaching this threshold, suggests that moderate agreement is attainable for certain distinctions, specifically for identifying

[14] For example, if annotator A1 marked turns 4 and 5 as part of the negotiation and annotator A2 marked turns 5, 6 and 7, A2's precision with respect to A1 is 1/3, recall is ½, and F1 is their harmonic mean, 0.4. Our agreement scores are symmetrical: any annotator can be treated as the gold standard.

Table 8 Krippendorff's alpha by individual label

	NON	DIN	WMN: Other	SIMN	Non-pursued	Without trigger	Reference/NE	Other kinds of clarification requests	Nothing
A1-R	0.58	0.20	0.53	0.53	0.52	**0.70**	0.49	0.47	0.63
A2-R	0.56	0.51	0.43	0.52	0.54	0.53	0.20	0.37	**0.65**
A3-R	0.64	0.43	0.41	0.55	0.53	**0.66**	0.40	0.51	**0.66**
A1-A2-A3	0.50	0.28	0.29	0.53	0.51	0.53	0.14	0.47	0.61
A1-A2-A3-R	0.55	0.35	0.37	0.53	0.52	0.58	0.27	0.46	0.63

complete or incomplete WMNs. For reference and to clarify the degree to which annotators agreed beyond chance, the observed agreement between all annotators and the expert annotation (A1-A2-A3-R) for the WMN-like categorization was 82%, compared to 53% expected by chance.

We also calculated agreement separately for each corpus (Table 9). All values that surpass or approach the moderate agreement threshold are in boldface. This analysis revealed a notable difference between Reddit and the spoken corpora. Agreement values are considerably higher for the spoken corpora, with A3 reaching 0.70 and 0.81 agreement with the reference on BNC and Switchboard. The other annotators also achieved values no lower than 0.64 for the WMN-like phenomena in these corpora. These results suggest that moderate agreement can be obtained for the general task of annotating WMN-like phenomena with the current training material. Further simplification of the guidelines and labelling scheme, along with more extensive training, could enhance consistency.

The lower agreement observed for the Reddit data partly explains the lower agreement values of DINs, which are almost exclusively found in that part of the corpus.[15] Reddit data is clearly more challenging to annotate, with its multiple intertwined threads and lengthy posts. In future work, we should consider modifying the data presentation for annotation, for example displaying only the relevant subthread instead of the entire thread and visually clarifying the tree structure.

A closer look at the confusion matrices (Figs. 2, 3 and 4 in the Appendix) reveals interesting insights into the nature of the differences in annotation. While each annotator exhibits specific tendencies, we observe two frequent patterns: (1) instances labeled as DINs in the reference are often annotated as NOTHING, and (2) the labels WMN: Other and NON are often confused with one another. Similarly, REFERENCE/NE is often annotated as NON or OTHER KINDS OF CLARIFICATION REQUESTS. Moreover, A1 labeled a large proportion of the sample as NOTHING. Examples 13 and 14 illustrate cases of disagreement between annotators. A larger training sample could have helped identify these patterns of inconsistency earlier, allowing us to provide more personalized feedback or refine the guidelines to clarify distinctions between the commonly confused classes. However, the limited availability of data for certain labels restricted the size and composition of the training sample.

S1: (...) The concept of a ***wiki*** requires universality to be useful (...)
S2: (...) A wiki does not in fact require universality to be useful. Specialist wikis abound (...) The original WikiWikiWeb site was the Portland Patterns Repository, an extremely specialist site about software design patterns, started in 1995 (six years before Wikipedia). And it isn't even a specialist *encyclopedia* - it includes *a couple decades worth* of long, rambling discussions on these and related topics. **Your claim as to what wikis are is incorrect about the very first wiki ever.**
S1: (...) *As for the definition of a wiki, I would argue that my definition is relatively accurate for *today*. I am relatively young and did not see the early days of the Internet, so my perspective is mostly from the modern era of the total dominance and mostly-trustworthiness of WP.*

[15] While it is possible that DINs are also inherently more difficult to agree upon, we observe a tendency for Reddit data to display lower agreement also in NONs and Nothings, which are common phenomena in the three corpora.

Table 9 Krippendorff's alpha for groups of phenomena, broken down by corpus

Annotators	BNC		Reddit		Switchboard	
	WMN	WMN-like	WMN	WMN-like	WMN	WMN-like
A1-R	0.63	**0.68**	0.35	0.48	0.62	**0.67**
A2-R	**0.68**	0.64	0.57	0.63	0.61	**0.65**
A3-R	**0.66**	**0.70**	0.57	0.58	**0.79**	**0.81**
A1-A2-A3	0.63	**0.68**	0.41	0.46	0.53	0.61
A1-A2-A3-R	**0.65**	**0.68**	0.46	0.52	0.60	**0.66**

Table 10 Krippendorff's alpha for the annotation of the type of meaning

Annotators	Krippendorff's alpha
A1-R	0.31 (98)
A2-R	0.31 (166)
A3-R	0.40 (142)
A1-A2	0.37 (108)
A1-A3	0.54 (93)
A2-A3	0.19 (146)
A1-A2-A3	0.37 (88)
A1-A2-A3-R	0.36 (83)

Table 11 IAA results on trigger, indicator and negotiation spans

	Trigger		Indicator	Negotiation
Annotators	Exact span match ratio	String match ratio (# comparisons)	% of coincidence (# comparisons)	F1-score (# comparisons)
A1-R	0.73	0.74 (222)	90% (261)	0.78 (134)
A2-R	0.69	0.77 (296)	87% (347)	0.73 (206)
A3-R	0.71	0.76 (273)	84% (315)	0.75 (186)
A1-A2	0.69	0.75 (233)	89% (276)	0.76 (144)
A1-A3	0.70	0.78 (220)	85% (265)	0.77 (131)
A2-A3	0.64	0.69 (296)	78% (350)	0.72 (191)

Example 13 An instance from Reddit annotated as a DIN by expert annotators and annotator A2, but as a NOTHING by annotators A1 and A3. Annotators may have viewed this as a discussion on topic, but there is an explicit disagreement and meta-linguistic shift on the definition of wiki.

S1: Some of the tracking control things and ***skidding control*** things for up north. The C D and the premium sound system.
S2: **Skidding control, you mean the antilock brake system?**
S1: *Yeah, it's kind of a traction control, I think they call it. It's not just antilock brake. I think that's already on most of them, but there is a further traction control*
S2: *Oh, this is a, probably suspension tied into the brakes*
(...)

Example 14 An instance from Switchboard annotated as a NON by expert annotators and annotator A3, but as a WMN: Other by annotators A1 and A2. While

the indicator is typical of a WMN: Other, the negotiation reveals a problem of understanding and, as explained in the annotation guidelines, such cases should be annotated as NON.

Finally, we present results of the agreement on the type of meaning in Table 10. The agreement values attested for this label range from 0.19 to 0.54. As we observed during the reference annotation process, this is quite a subjective notion that is hard to agree upon. In future annotation efforts including the type of meaning, it may be necessary to provide more training dedicated to this notion.

6.2.2 Span agreement

In Table 11, we present results of span-related agreement. Once annotators agree that an instance contains a phenomenon requiring a specific kind of span, the agreement on the location of these spans shows higher consistency. Match ratios for trigger spans are between 0.64 and 0.78, indicating a moderate but acceptable agreement. For indicators, agreement is even higher, with indicators being located in the same utterance as the reference between 84 and 90% of the time. We expected indicators to show a high agreement, because they tend to be the most easily identifiable part of a WMN. Finally, F1-scores for negotiation spans are all above 0.72, indicating more than acceptable agreement.

Agreement on spans shows less variation across annotators compared to other measures, and agreement values with the reference annotations are comparable to those observed between pairs of annotators. This suggests that the guidelines for span identification are mostly clear and that, despite some degree of disagreement and subjectivity, identifying spans is a relatively straightforward task.

7 Conclusion

We have presented the NeWMe corpus, the first corpus dedicated to the analysis of Word Meaning Negotiation (WMN). The NeWMe corpus contains manual annotations for a total of 404 WMN sequences as well as other related phenomena, including incomplete WMNs and distractors. The corpus includes interactions recorded or collected between approximately 2000 and 2015. Given this temporal range, the NeWMe corpus is not designed to evaluate the contemporary meaning of specific words. Rather, our goal has been to provide a gold-standard resource for examining how word meaning is negotiated in interaction. Since discourse-level strategies and meta-linguistic engagement can be studied regardless of the time period, the corpus offers a foundation to support future research on the processes by which participants clarify, challenge, and negotiate meaning. To our knowledge, the NeWMe corpus

is the first resource to systematically identify and annotate meta-linguistic shifts in conversation, focusing specifically on word meaning negotiation.

Our analysis of the annotated data revealed several interesting tendencies regarding the prevalence of each phenomenon across different forms of interaction and activity types. Our annotation of the type of meaning showed that, while most WMNs center on situated meaning, DINs tend to discuss potential meaning more often than NONs. This distinction underscores the different functions of these negotiation types: NONs primarily aim at resolving immediate comprehension issues, while DINs often involve broader debates on word definitions. We also found that failed and abandoned attempts at initiating WMN (Non-pursued sequences) and negotiations arising from disagreements about word meaning are much more common in asynchronous online interaction than in spoken interaction. This pattern reflects the unique characteristics of each type of interaction. In asynchronous platforms like online forums, participants can take their time to respond or choose not to respond at all, making it easier for disagreements to drag on or for clarification attempts to be ignored. In contrast, spoken interactions occur in real-time, where the flow of conversation relies on immediate turn-taking and sequential progression. However, it remains unclear to what extent these differences are driven by the type of interaction activity rather than the mode of the interaction itself. The online data in our corpus mainly consists of open-ended discussions, while the spoken data includes a greater variety of interactional contexts, including interviews, meetings and informal discussions. Open-ended discussions may naturally involve more disagreement, contributing to the higher prevalence of DINs in the online data. Future research could aim at disentangling the role of activity type from the interaction mode in shaping WMN frequency and characteristics.

While the agreement in the reference annotation is high, our IAA analysis with three external annotators revealed areas of improvement. The IAA results broken down by corpus and label show that an acceptable level of agreement is attainable despite the subjectivity of the task, but that (1) the annotation setup and labels could be simplified, (2) written threaded online discussions such as those found in Reddit should be presented in a more streamlined format, and (3) more extensive and targeted annotator training should be provided to enhance consistency.

We also carried out an analysis of our regular expression-based methodology for candidate selection, which served to evaluate its efficiency and point to directions of improvement for future work. This analysis identified expressions that are highly indicative of a WMN, as well as others that require additional filtering or refinement to improve precision and reduce irrelevant matches. For example, while expressions such as “what do you mean by X?” had a high precision (40.9%), broader expressions like “what is” and “this is not X” had much lower precision (4.9% and 3.1%, respectively). Refining these regexes and supplementing them with machine learning-based filtering could improve both precision and recall in future iterations of corpus development, reducing annotation workload and enhancing the overall efficiency of data

collection. We also observed differences depending on the type of WMN: indicators including "you mean" were more often involved in NONs, while expressions invoking the definition of a word (e.g., "can you define") were mostly used in DINs.

The current approach does not allow for determining the true prevalence of WMNs in the selected corpora, as the analysis is limited to interactions captured by the regular expressions. The number of WMN instances using other expressions that may have been missed remains unknown, emphasizing the need for further methodological refinement and exploration. While our method successfully identified 404 WMNs, its recall remains unknown since we do not have a measure of the actual prevalence of WMNs in the corpora. Estimating recall would require manually reading through a large portion of the corpus, which is highly impractical for a phenomenon that appears relatively rare: we found WMNs in 11% of all conversations considered, which can be regarded as a lower bound of its frequency in the data used.

In future work, we plan to expand this corpus by including additional languages, conversational settings, and modalities. This extension will incorporate the lessons learned from this study, complementing regular expression searches with semi-automatic corpus expansion methods informed by our manual annotations. For example, we may find additional ways of expressing an indicator that we had not previously considered by looking into indicators that were annotated in the vicinity of a regex match. Additionally, we aim to explore machine learning approaches to enhance the identification of WMNs, leveraging the annotated data as training material; and the use of large language models as an alternative retrieval method. Future work could also explore dialogue acts and strategies for contributing to ongoing WMN sequences, drawing on Myrendal's (2015) taxonomy of WMN-related dialogue acts, which includes mechanisms such as explicification, exemplification, and contrast.

The standoff annotations are made available[16] to the research community, accompanied by tools and instructions for retrieving the corpora and inspecting annotated interactions. We also make available the complete annotation guidelines, enabling further research on WMN and related phenomena such as conversational repair and clarification requests. Furthermore, the annotated instances are freely browsable online, providing an accessible platform for researchers and practitioners to explore the corpus and draw insights from its rich dataset. We also provide the actual regular expressions used to match the relevant utterances in the corpora and the relevant scripts used for retrieving and analyzing data. By making these resources widely available, we hope to advance the empirical and theoretical study of WMNs and support the development of practical applications, such as adaptive dialogue systems capable of navigating and resolving potential semantic misalignments in real-time interaction.

Appendix 1

See Figs. 1, 2, 3, 4.

[16] https://github.com/gu-wmn/NeWMe/

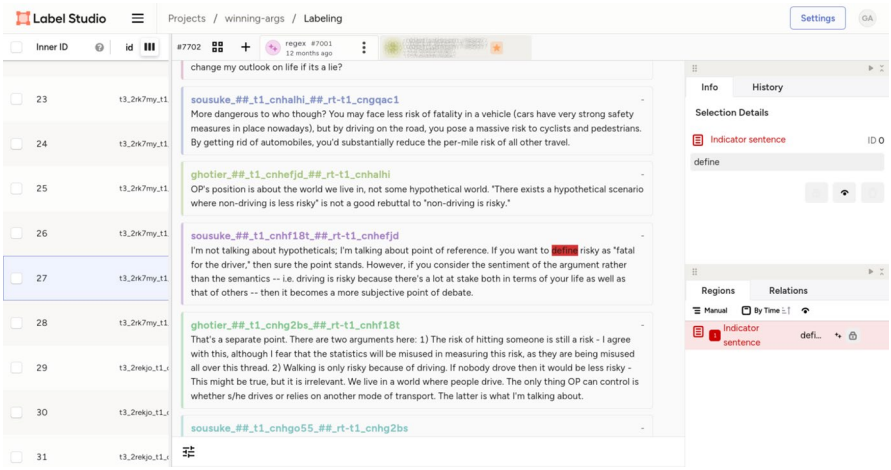

Fig. 1 The LabelStudio annotation interface

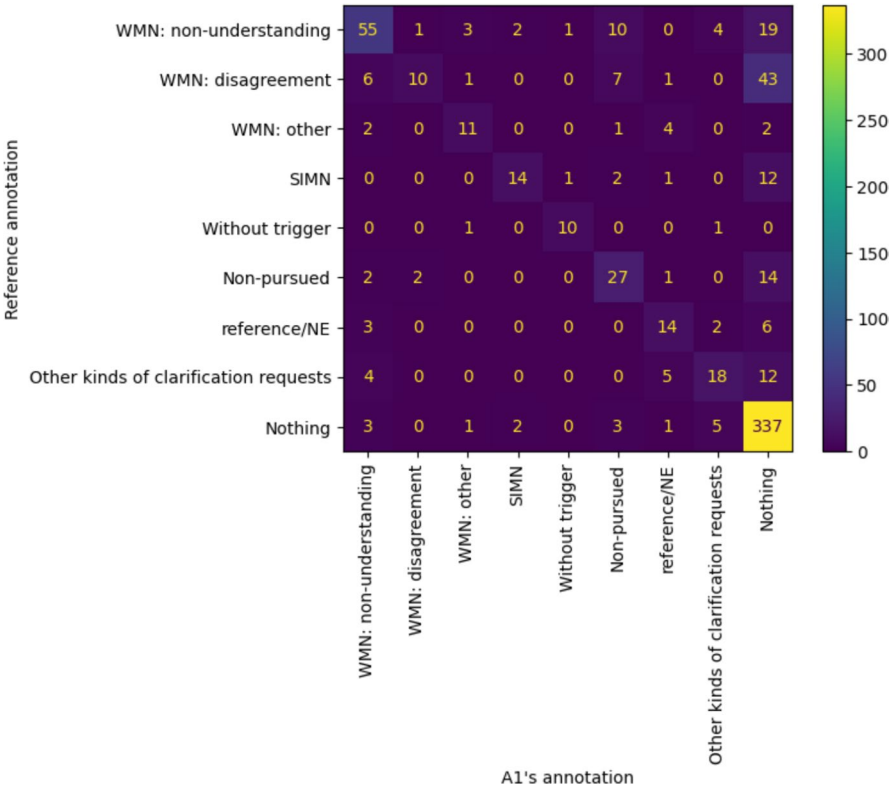

Fig. 2 Confusion matrix between the reference annotation and A1's annotation

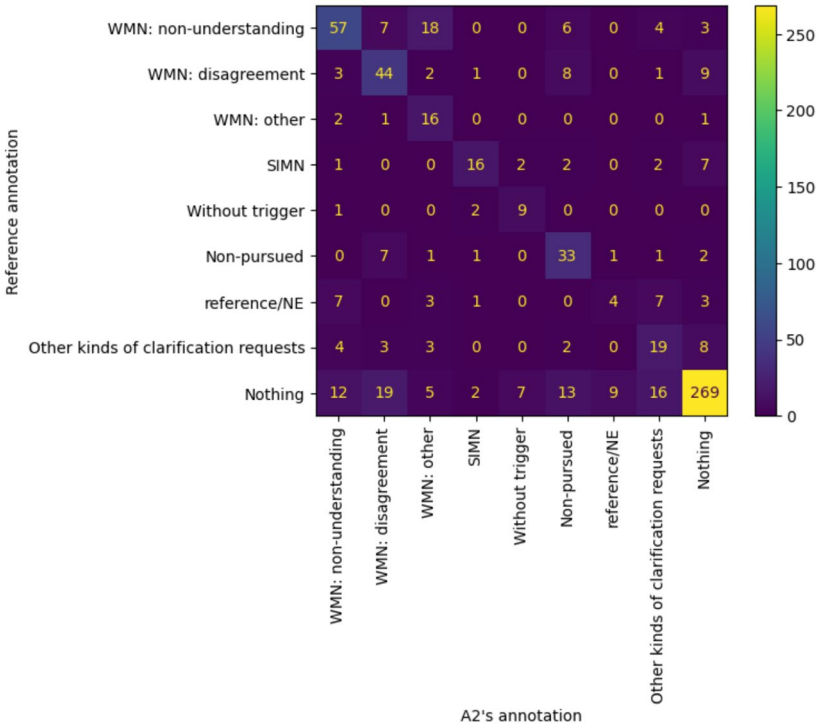

Fig. 3 Confusion matrix between the reference annotation and A2's annotation

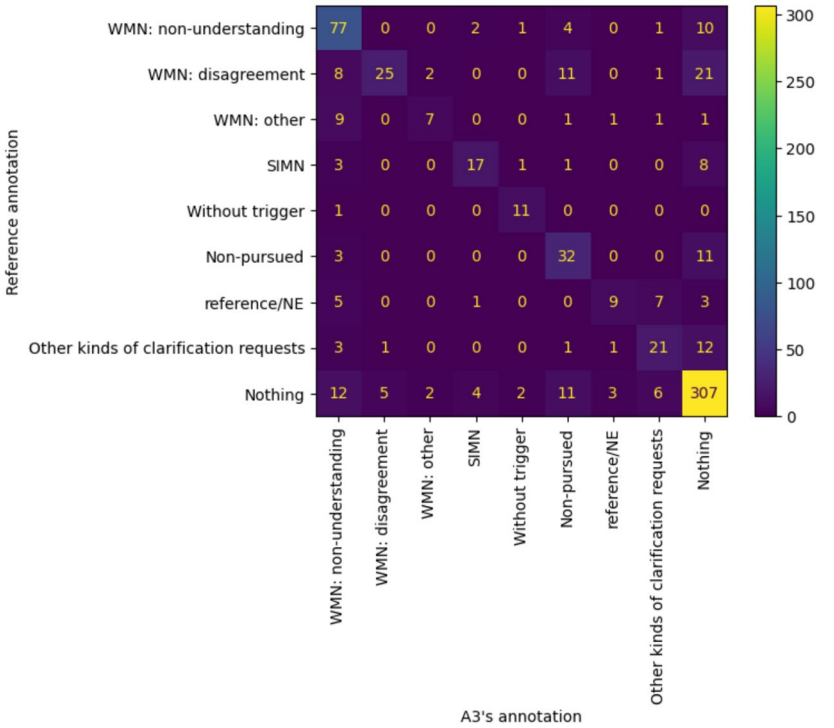

Fig. 4 Confusion matrix between the reference annotation and A3's annotation

See Tables 12 and 13.

Table 12 Complete table of regex success by corpus

Regex name	BNC			Switchboard			Reddit		
	# matches	# WMN	% WMN	# matches	# WMN	% WMN	# matches	# WMN	% WMN
you mean	424	61	14.4	91	19	20.9	319	92	28.8
can you define	263	7	2.7	16	1	6.2	574	141	24.6
definition of	95	4	4.2	2	0	0.0	384	117	30.5
what is	812	35	4.3	439	10	2.3	233	20	8.6
what do you mean by	61	23	37.7	2	2	100.0	91	38	41.8
this is not X	809	8	1.0	244	0	0.0	876	51	5.8
is that a/the	219	3	1.4	263	1	0.4	470	33	7.0
this isn't X	264	1	0.4	68	2	2.9	541	33	6.1
[repetition pattern]	888	26	2.9	113	4	3.5	5	0	0.0
what X means	202	3	1.5	11	0	0.0	133	25	18.8
the term	90	4	4.4	11	0	0.0	56	16	28.6
in what way/sense	43	3	7.0	5	0	0.0	25	8	32.0
what does X mean	96	9	9.4	0	0	0.0	12	2	16.7
are you talking about	18	0	0.0	16	3	18.8	18	4	22.2
what is the difference between	49	5	10.2	5	2	40.0	16	0	0.0
isn't this…?	33	0	0.0	0	0	0.0	45	7	15.6
what sort/kind of	100	5	5.0	25	1	4.0	7	0	0.0
elaborate	41	2	4.9	13	0	0.0	65	4	6.2
the meaning of	15	1	6.7	1	0	0.0	35	4	11.4
is this not…?	9	0	0.0	4	0	0.0	22	4	18.2
are we talking	26	0	0.0	3	1	33.3	9	2	22.2
do you know what X is	26	1	3.8	9	0	0.0	8	1	12.5
is this what you mean	12	1	8.3	1	0	0.0	4	1	25.0
word X means	18	1	5.6	0	0	0.0	10	1	10.0
have you heard of	20	0	0.0	10	1	10.0	10	0	0.0
is that what X is	11	0	0.0	3	0	0.0	1	1	100.0

Table 12 (continued)

Regex name	BNC			Switchboard			Reddit		
	# matches	# WMN	% WMN	# matches	# WMN	% WMN	# matches	# WMN	% WMN
does X count as Y	1	0	0.0	0	0	0.0	5	1	20.0
how do you mean that	0	0	0.0	2	0	0.0	0	0	0.0
that would mean	19	0	0.0	3	0	0.0	26	0	0.0
is that what that/this is	0	0	0.0	1	0	0.0	0	0	0.0

Table 13 Complete table of regex success in relation to WMN type

Regex name	# WMNs	# NONs	# DINs	# Other	% NONs	% DINs	% Other
you mean	172	107	45	20	62.2	26.2	11.6
can you define	149	42	107	0	28.2	71.8	0.0
definition of	121	11	110	0	9.1	90.9	0.0
what is	65	46	18	1	70.8	27.7	1.5
what do you mean by	63	46	16	1	73.0	25.4	1.6
this is not X	59	3	55	1	5.1	93.2	1.7
is that a/the	37	9	28	0	24.3	75.7	0.0
this isn't X	36	7	29	0	19.4	80.6	0.0
[repetition pattern]	30	29	0	1	96.7	0.00	3.3
what X means	28	5	23	0	17.9	82.1	0.0
the term	20	4	16	0	20.0	80.0	0.0
in what way/sense	11	3	8	0	27.3	72.7	0.0
what does X mean	11	10	1	0	90.9	9.1	0.0
are you talking about	7	4	1	2	57.1	14.3	28.6
what is the difference between	7	6	1	0	85.7	14.3	0.0
isn't this…?	7	3	4	0	42.9	57.1	0.0
what sort/kind of	6	6	0	0	100.0	0.0	0.0
elaborate	6	5	1	0	83.3	16.7	0.0
the meaning of	5	1	4	0	20.0	80.0	0.0
is this not…?	4	1	3	0	25.0	75.0	0.0
are we talking	3	3	0	0	100.0	0.0	0.0
do you know what X is	2	1	1	0	50.0	50.0	0.0
is this what you mean	2	1	1	0	50.0	50.0	0.0
word X means	2	0	2	0	0.0	100.0	0.0
have you heard of	1	1	0	0	100.0	0.0	0.0
is that what X is	1	0	1	0	0.0	100.0	0.0
does X count as Y	1	0	1	0	0.0	100.0	0.0
how do you mean that	0	0	0	0	–	–	–
that would mean	0	0	0	0	–	–	–
is that what that/this is	0	0	0	0	–	–	–
depends on what you mean	0	0	0	0	–	–	–

Acknowledgements This work was supported by the Swedish Research Council (VR) grant 2022-02125 Not Just Semantics: Word Meaning Negotiation in Social Media and Spoken Interaction, and by state funding managed by the Agence Nationale de la Recherche under the France 2030 program, with reference "ANR-23-IACL-0008". We want to thank Matthieu Labeau and William Noble for their contributions to the discussions, Anh Ngo Ha for her feedback on the annotation guidelines, and Kaj Ailomaa for his work on the website infrastructure and GitHub repository, as well as the three annotators for their efforts in the annotation process.

Author contributions A.G.S (Aina Garí Soler) and J.M. (Jenny Myrendal) led the effort and are the main authors. All authors collaboratively contributed to the conceptualization, methodology, analysis and discussion of the results. A.G.S. and J.M. performed manual annotation, writing of the guidelines, annotator training and the writing of the original draft. A.G.S. led data collection and formal data analysis. J.M. led annotator recruitment. C.C. (Chloé Clavel) and S.L. (Staffan Larsson) played central roles in framing the findings within broader theoretical and methodological contexts and reviewing and improving the manuscript.

Funding Open access funding provided by Télécom Paris. This work was supported by the Swedish Research Council (VR) Grant 2022-02125 Not Just Semantics: Word Meaning Negotiation in Social Media and Spoken Interaction, and by state funding managed by the Agence Nationale de la Recherche under the France 2030 program, with reference "ANR-23-IACL-0008".

Data availability The standoff annotations are made available to the research community, accompanied by tools and instructions for retrieving the corpora and inspecting annotated interactions. We also make available the complete annotation guidelines, enabling further research on WMN and related phenomena such as conversational repair and clarification requests. Furthermore, the annotated instances are freely browsable online, providing an accessible platform for researchers and practitioners to explore the corpus and draw insights from its rich dataset. We also provide the actual regular expressions used to match the relevant utterances in the corpora and the relevant scripts used for retrieving and analyzing data. All of this is available in the project's GitHub repository: https://github.com/gu-wmn/NeWMe/.

Declarations

Conflict of interest The authors have no financial or proprietary interests in any material discussed in this article.

References

Babcock, M. J., Ta, V. P., & Ickes, W. (2014). Latent semantic similarity and language style matching in initial dyadic interactions. *Journal of Language and Social Psychology, 33*(3), 77–88.

Barnidge, M. (2017). Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political Communication, 34*(2), 302–321. https://doi.org/10.1080/10584609.2016.1235639

BNC Consortium. (2007). *The British National Corpus, XML Edition*. Oxford Text Archive. http://hdl.handle.net/20.500.14106/2554

boyd, dm, & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-mediated Communication, 13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13–B25. https://doi.org/10.1016/S0010-0277(99)00081-5

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482–1493.

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020, July). ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue.*

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language, 49*(2), 214–230.

Drew, P. (1997). 'Open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics, 28*(1), 69–101.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-mediated Communication, 12*(4), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x

Garí Soler, A., Labeau, M., & Clavel, C. (2023). Measuring lexico-semantic alignment in debates with contextualized word representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)* (pp. 50–63). Toronto, Canada. https://doi.org/10.18653/v1/2023.sicon-1.6

Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall.

Heritage, J. (1984). *Garfinkel and ethnomethodology*. Polity Press.

Hough, J., & Purver, M. (2013). Modelling expectations in the self-repair processing of annot-, um, listeners. In R. Fernández & A. Isard (Eds.), *Proceedings of the 17th SemDial workshop on the semantics and pragmatics of dialogue* (pp. 92–101). Amsterdam, Netherlands, 16–18 December 2013.

Karafoti, E. (2021). Negotiating preferred norms in requests and offers: Is the (dis)preferred answer so obviously (im)polite? *Journal of Pragmatics, 173*, 134–147. https://doi.org/10.1016/j.pragma.2020.12.012

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Sage.

Larsson, S., & Myrendal, J. (2017a). Towards dialogue acts and updates for semantic coordination. In *Proceedings of the workshop on formal approaches to the dynamics of linguistic interaction (FADLI 2017).*

Larsson, S., & Myrendal, J. (2017b). Dialogue acts and updates for semantic coordination. In V. Petukhova & Y. Tian (Eds.), *Proceedings of the 21st workshop on the semantics and pragmatics of dialogue, Saarbrücken, 15–17 August 2017*.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of research on language acquisition: Second language acquisition* (Vol. 2, pp. 413–468). Academic Press.

Myrendal, J. (2015). *Word meaning negotiation in online discussion forum communication* (Doctoral dissertation). University of Gothenburg.

Myrendal, J. (2019). Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse Studies, 21*(3), 1–23.

Myrendal, J. (2025). Repair of claimed non-understanding of word meaning in online discussion forum interaction. *Dialogue & Discourse, 16*(1), 91–113. https://doi.org/10.5210/dad.2025.104

Myrendal, J., & Larsson, S. (2025). Semantic conflict in online discussions: Negotiating the meaning of 'lying.' *Journal of Language Aggression and Conflict*. https://doi.org/10.1075/jlac.00133.myr

Noble, B., Viloria, K., Larsson, S., & Sayeed, A. (2021). What do you mean by negotiation? Annotating social media discussions about word meaning. In *Proceedings of the 25th workshop on the semantics and pragmatics of dialogue - full papers.*

Noble, B., Larsson, S., & Myrendal, J. (2025). Misunderstanding the concrete, disagreeing about the abstract: A closer look at word meaning negotiation triggers. In *SemDial 2025—The 29th workshop on the semantics and pragmatics of dialogue*.

Norén, K., & Linell, P. (2007). Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics, 17*(3), 387–416. https://doi.org/10.1075/prag.17.3.03nor

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–226.

Pitzl, M.-L. (2005). Non-understanding in English as lingua franca: Examples from a business context. *Vienna English Working Papers, 14*(2), 50–71.
Pomerantz, A., & Heritage, J. (2013). Preference. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 210–228). Wiley-Blackwell.
Purver, M. (2004). *The theory and use of clarification requests in dialogue* (Ph.D. thesis). University of London.
Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 175–190). Seattle, United States. Association for Computational Linguistics.
Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.
Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge University Press.
Schober, M. F. (2005). Conceptual alignment in conversation. In S. T. Fiske, H. R. Markus, & P. A. Glick (Eds.), *Other minds: How humans bridge the divide between self and others* (pp. 239–252). Guilford Press.
Stivers, T., & Robinson, J. D. (2006). A preference for progressivity in interaction. *Language in Society, 35*(3), 367–392.
Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences, 20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007
Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics, 26*(3), 339–373.
Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613–624).
Varonis, E. M., & Gass, S. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics, 6*(1), 71–90.
Vasseur, M.-T., Broeder, P., & Roberts, C. (1996). Managing understanding from a minority perspective. In K. Bremer, C. Roberts, M.-T. Vasseur, M. Simonot, & P. Broeder (Eds.), *Achieving understanding: Discourse in intercultural encounters* (pp. 65–108). Longman.

Authors and Affiliations

Aina Garí Soler[1,2] · Jenny Myrendal[3,4] · Chloé Clavel[1,2] · Staffan Larsson[3]

✉ Aina Garí Soler
aina.gari-soler@inria.fr

[1] INRIA Paris, Paris, France

[2] LTCI, Télécom-Paris, Institut Polytechnique de Paris, Palaiseau, France

[3] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden

[4] Department of Education, Communication and Learning, University of Gothenburg, Gothenburg, Sweden