# What do you mean by *negotiation*? Annotating social media discussions about word meaning

**Bill Noble** and **Kate Viloria** and **Staffan Larsson** and **Asad Sayeed**
Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg
bill.noble@gu.se
kateviloria@outlook.com
staffan.larsson@ling.gu.se
asad.sayeed@gu.se

## Abstract

We present a formalisation and annotation protocol for *word meaning negotiation* (WMN), a conversational routine in which speakers explicitly discuss the meaning of a word or phrase. WMN is formalised as an interaction game with a shared game board and rules for subsequent contributions, as well as a semantic update function based on the state of the game board. We develop an annotation schema based on this formalisation and present the results of annotating 150 Twitter conversations as WMNs.

## 1 Introduction

Meaningful dialogue requires some degree of alignment between participants' lexico-semantic resources. When misalignments are discovered, participants may choose to explicitly engage with the discrepancy in a metalinguistic discussion where the meaning of a misaligned word or phrase is at issue. These discussions—termed *word meaning negotiations* (WMN)—exhibit a certain structure, which we attempt to characterise and put to use by annotating WMNs collected from Twitter.

The opportunity for a WMN arises whenever a dialogue participant finds that they disagree with—or do not understand—what another speaker meant by a certain *trigger word* or phrase. They may ignore the discrepancy or silently deal with on their own (Larsson, 2010), or they may *indicate* it to their interlocutor (perhaps in the form of a clarification request). If the interlocutor responds to the indicator, a WMN has been initiated. As the WMN progresses, participants may propose, accept, reject, or raise the question of particular semantic relations between the word that triggered the WMN and other entities, which we refer to as *anchors*.

We start by discussing previous work on WMN that underpins this contribution (Section 2). Then, we develop the formal model of WMN, including a semantic update rule that can be integrated in a game board model of dialogue (Section 3). After that, we introduce an annotation schema, based on our WMN model, and present the results of an annotation study using that schema (Section 4). Finally, we discuss insights into the phenomenon of WMN resulting from the annotation study and suggest avenues for future work (Section 5).

## 2 Background and Related Work

There is surprisingly little work on word meaning negotiation as such. WMNs, in the form of corrective feedback, have been studied as an aspect of first language acquisition (Clark, 2007). There has also been work that teaches artificial agents the meaning of novel terms based on defintions and grounded perceptual examples (Mohan et al., 2012; Krause et al., 2014). WMNs have also been studied in conversations between non-native language learners (Varonis and Gass, 1985; Long, 1996). Myrendal (2015, 2019) has taken a more in-depth look at WMNs between adult speakers, focusing on conversations in Swedish online discussion forums.

The model and annotation scheme we develop in this work builds on the structural model of Varonis and Gass (1985) and the classificatory schemas of Myrendal (2015, 2019). The semantic update fuction we define in Section 3.4 extends the dialogue acts proposed by Larsson and Myrendal (2017). We discuss this foundation in more depth below.

**TIR model** In the Trigger-Indicator-Response model, when an interlocutor recognizes a non-understanding and chooses to address it overtly, the discourse enters a "subroutine" in which participants attempt to repair the non-understanding and align their semantic common ground. These subroutines are embedded in the regular linear flow of dialogue in such a way that the current line of con-

*Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.*

versation is suspended Furthermore, WMNs may be nested if, in the course of resolving one non-understanding, another non-understanding occurs and is indicated by one of the participants.

A WMN has three key elements:

**Trigger** – an utterance by a speaker, $S_1$, that contains a lexical item resulting in non-understanding by another participant, $S_2$.

**Indicator** – an utterance in which $S_2$ explicitly indicates their non-understanding of the trigger.

**Response** – an utterance in which $S_1$ overtly acknowledges the non-understanding.

A trigger can occur at any point in a dialogue (e.g., in a question *or* in a response). The non-understanding is only made part of the common ground once it has been indicated by $S_2$—thus, the trigger can only be identified retrospectively, with respect to its indicator. Likewise, the response refers back to the indicator: it may attempt to rectify the non-understanding, or merely acknowledge that a discrepancy was indicated.

Although the T-I-R model was developed for WMNs in a language learning context, Myrendal (2015) found it to be a good model for the initiation of WMNs in discussion forums as well.

**Non-understanding vs. disagreement** Myrendal (2015) categorises WMNs as those resulting from *misunderstanding* (NON), when one dialogue participant doesn't understand the meaning of a word uttered by another participant, in the context in which it was used, or *disagreement* (DIN), when a participant disagrees with how someone else used a word,(Myrendal, 2015). NONs are generally initiated with a *metalinguistic clarification request*, whereas DINs are initiated with a *metalinguistic objection*.

**WMN dialogue acts** Myrendal (2019) inventories types of WMN contributions, including *generic* and *specific explicifications*[1] (which we refer to as *partial definitions*), *exemplification*, *contrasting*, *metalinguistic objections* (which can be used in an ongoing WMN, as well as to initiate one), and *endorsement* (of a using a particular word in a given context).

Larsson and Myrendal (2017) propose dialogue acts based on these contribution types, and propose semantic update functions for exemplification partial definition and contrasting, that apply to the meaning of the trigger word, in the event that the dialogue act is grounded. In this paper, we expand on that work by using the act-level update functions to define an update that takes the entire WMN into account.

[1](see also Ludlow, 2014).

## 3 Formal model

The model presented in this section has a dual purpose. First, it affords the precise formulation of hypotheses about WMNs (in general or in a particular domain) that can be tested in terms of the model. Second, the model itself implies a certain structure to the phenomenon of WMNs which may, to a greater or lesser degree, capture what is observed. As is often the case, these two roles are not entirely separable: What is expressible in the model affects the hypotheses that can be tested; How well the model aligns empirically with the phenomenon it seeks to describe affects the reliability of the conclusions one can draw.

In addition to the descriptive goal, we want the model to support a semantic *update function* that computes the change in shared lexical resources resulting from a WMN (section 3.4). The rule we define builds on the work of (Larsson and Myrendal, 2017), taking their dialogue act-specific rules and extending them to operate over a whole WMN.

Our model of word meaning negotiation depends on the notion of semantic *anchors* and speaker commitments to semantic *relations* between those anchors. This is motivated by the intuition that when speakers discuss the meaning of a word, they do so by triangulating it in reference to other points (or regions) of semantic space. In a successful WMN, the meaning of the word in question is "anchored" by the participants as a result of joint commitment to relations between the word and reference points (i.e., *anchors*) that *are* grounded.

When the project of aligning on meaning has started, it is not uncommon to discover that further discrepancies exist; that is, it can be that some of the anchors introduced to negotiate the meaning of the trigger word are themselves lacking semantic common ground (as in Varonis and Gass, 1985). This shouldn't be surprising: First of all, once a WMN has begun, discrepancies that might have gone unnoticed or un-remarked-upon are suddenly difficult to ignore. Furthermore, new anchors are introduced precisely *because* one of the participants thinks they have an elucidating relation to

the trigger. Where one semantic misalignment exists, misalignment on related terms may be lying in wait. What makes something eligible as an anchor is not that its *meaning* is common ground and fully specified, but that it can be grounded as a shared *discourse referent*, available for participants invoke anaphorically (or by name or description) and put in relation to other anchors as well as to the trigger.

We represent a word meaning negotiation, between a set of speakers $S$ taking place over $N$ turns, as sequence of tuples:

$$\mathsf{WMN} = \langle s_i, A_i, R_i \rangle_{i \leq N} \tag{1}$$

where $s_i$ is the speaker at turn $i$, $A_i$ is the set of anchors introduced in that turn (we let $t \in A_0$ be the trigger), and $R_i$ is the set of relations between anchors that $s_i$ publicly commits (Asher and Lascarides, 2008) to during that turn.

### 3.1 Anchors

Once introduced, anchors are available for the remainder of the WMN, accessible by co-referring expressions, including anaphora. Thus, the set of common ground anchors at turn $i$ is defined as the union of anchors introduced so far:

$$\boldsymbol{A}_i = \bigcup_{j \leq i} A_i \tag{2}$$

We let $[\![a]\!]$ denote the meaning of $a$, given the context of the dialogue and the semantic common ground of the speakers, without yet considering any updates resulting from the WMN.[2]

### 3.2 Semantic relations

Word meaning negotiation depends on a commonly understood set of possible semantic *relation types* between anchors, $\mathcal{R}$. In the remainder of the formalisation and in the annotation study (Section 4), we assume two semantic relations, *example* and *partial definition*:

$$\mathcal{R} = \{\mathsf{Exa}, \mathsf{Def}\} \tag{3}$$

We also make use of a set of *polarities*:

$$\mathcal{O} = \{+, -, ?\} \tag{4}$$

Polarity correspond to an attitude (or commitment) that speakers may express towards a given relation between two anchors. This set of polarities indicate whether a relation holds ($+$) or its converse holds ($-$), or if the matter is in question ($?$).

In the model, $R_i \subseteq \mathcal{R} \times \mathcal{O} \times \mathbf{A}_i \times \mathbf{A}_i$ is a set of semantic relations. We will write $\mathsf{R}^o(a, b)$ for $\langle \mathsf{R}, v, a, b \rangle$. For example, $\mathsf{Def}^+(a, b) \in R_i$ means that speaker $s_i$ has publicly committed to $a$ as a (positive) partial definition of $b$.

Given $\mathsf{WMN}$, we can compute a speaker's current commitments. For a pair of anchors $(a, b)$ and relation $\mathsf{R}$, we consider the speaker to be committed to the most recent polarity that has been part of their public commitments. Formally, this is defined as follows:

$$\boldsymbol{R}_{s,0} = \begin{cases} R_0 & \text{if } s = s_0 \\ \varnothing & \text{otherwise} \end{cases} \tag{5}$$

and

$$\boldsymbol{R}_{s,i+1} = \begin{cases} \boldsymbol{R}'_{s,i} \cup R_{s,i+1} & \text{if } s = s_i \\ \boldsymbol{R}_{s,i} & \text{otherwise} \end{cases} \tag{6}$$

where

$$\boldsymbol{R}'_{s,i} = \{\mathsf{R}^o(a, b) \in \boldsymbol{R}_{s,i} \mid \neg \exists o'.\, \mathsf{R}^{o'}(a, b) \in R_s\} \tag{7}$$

Finally, we define the common ground relations at turn $i$ as those relations to which all speakers have publicly committed:

$$\boldsymbol{R}_i = \bigcap_{s \in S} \boldsymbol{R}_{s,i} \tag{8}$$

### 3.3 Interaction rules

Now that we have a structure for representing the state of a WMN at each turn and a way to compute what is common ground based on the history of those states, we characterise the rules of the WMN as an interaction game.

Formally, there are very few conditions on what $A_i$ and $R_i$ can include. Any number of anchors can be introduced in a turn, although practically the number is usually quite small (see Section 4.4). The main restriction on $R_i$ is that it must not result in a cycle in $s_i$'s public commitments; that is, $\{(a, b) \mid \mathsf{R}^o(a, b) \in \boldsymbol{R}_{i,s_i}\}$ must not contain a cycle. This means that $\boldsymbol{R}_{i,s}$, considered as a labeled directed graph, is acyclic, a condition that is necessary for the semantic update function (Section 3.4) to be well-defined. Intuitively it would be very

[2] Note that this interpretation, as with the negotiated meaning defined in Section 3.4, may be different for different speakers, since speakers can of course be wrong about what is common ground.

strange for speakers to ground such a cycle for exactly that reason—indeed we did not see any such cycles in speaker commitments (let alone grounded cycles) in our annotation study, although the annotation protocol would have allowed it. There are three ways of contributing to $R_i$:

**Propose (or raise) a relation** For any two anchors in $\boldsymbol{A}_i$, the speaker either proposes a relation between them ($o \in \{+, -\}$) or poses the question of their relation without asserting anything one way or the other ($o \in \{?\}$).

**Ground a relation** The speaker makes some indication of their stance (or negative grounding) regarding a relation that another speaker has just committed to. For some $\mathsf{R}^o(a, b) \in R_{i-1}$, $\mathsf{R}^{o'}(a, b) \in R_i$, where $o, o' \neq ?$. If $o = o'$, then it is *positive grounding*, otherwise it is *negative grounding*.

Positive grounding can be accomplished more or less implicitly, though what counts as grounding may depend on the WMN type (NON or DIN), as well as other factors such as the medium of the dialogue and social context.

**Answer a question** Finally, for $\mathsf{R}^?(a, b) \in R_{i-1}$, $s_i$ can add $\mathsf{R}^o(a, b)$ to $R_i$ for any $o \neq ?$ by answering the question posed by $s_i$. Note that *grounding a relation* and *answering a question* don't formally add to the possible elements of $R_i$ beyond *posing a relation*, but we characterise them separately because they usually take the form of grounding statements or polar answers which don't include explicit co-reference to an anchor. For that reason, we also annotate them differently (Section 4.2).

### 3.4 Semantic update

Our goal is to define a semantic update function that takes WMN as input. We define update functions that apply to the meaning of an anchor, based on a relation with another anchor, if that relation is grounded. Then, we recursively define the update for a whole WMN based on those functions in a straightforward way:

For $a \in \boldsymbol{A}_N$, let

$$\{\mathsf{R}_1^{o_1}(b_1, a), ..., \mathsf{R}_n^{o_n}(b_n, a)\} \subseteq \boldsymbol{R}_N$$

be the common round relations anchoring $a$ at turn $N$. Then the semantic update given by WMN for $a$ is defined as:

$$\Delta(a) = [I(\mathsf{R}_1, o_1, \Delta(b_1)) \circ ... \circ I(\mathsf{R}_n, o_n, \Delta(b_n))](\llbracket a \rrbracket) \quad (9)$$

Here, $I$ is the interpretation of R (we assume that for a semantic relation to be common ground implies the existence of an update function):[3]

$$I = \begin{cases} \lambda x.\epsilon^o(b, x) \text{ if } \mathsf{R} = \mathsf{Exa} \\ \lambda x.\delta^o(b, x) \text{ if } \mathsf{R} = \mathsf{Def} \end{cases} \quad (10)$$

In essence, $\Delta$, as defined in (9) applies the update implied by the semantic relations recursively on $\mathbf{R}_N$ in a straightforward way: the updated meaning of an anchor is computed by sequentially applying each its grounded relations to other anchors, with the caveat that each of those anchors should first have *their* meaning updated, if they were also negotiated as part of the WMN.

## 4 Annotation study

### 4.1 Data

We collected exchanges on Twitter that, based on search heuristics, were likely to involve WMN. In particular, we used the Twitter filtered stream API to find tweets that were in reply to another tweet and that used the indicator phrase *what do you mean by*.[4] This heuristic method is based on that of Myrendal (2015), who used similar phrases in Swedish to build a corpus of WMNs from online discussion forums. The search resulted in a total of 1783 candidate indicator tweets, collected over a 24-hour period (May 5–6, 2021).

After 48 hours (to wait for replies), we used the Twitter search API to collect the rest of the thread, retrieving tweets both upwards and downwards in the reply chain. Since the reply structure on Twitter is a tree (each tweet can be *in reply to* at most one other tweet, but can *have* multiple replies), retrieving the upwards context is easy—we just followed the replies up to the root of the thread. For the downward search (replies to the indicator), we initially look for a reply from the author that the indicator was a reply to, alternating back and forth between these two users for further replies and taking the first reply in case there were multiple.[5] This resulted in 671 threads with at least one reply after the candidate indicator (38% of threads), of which we randomly sampled 150 for annotation.

[3] We let $\epsilon^+$, $\epsilon^-$, $\delta^+$, and $\delta^-$ be as defined in Larsson and Myrendal (2017).

[4] We used a regular expression to allow for some variation in the exact wording (see supplementary materials for details).

[5] This is a somewhat brittle heuristic that could improved upon. For example, it breaks if a user makes a "double reply" or if the conversation is between more than two users.

### 4.2 Annotation protocol

The annotation protocol, which was developed over a series of pilot studies, aims to be comprehensible for annotators with no linguistic background (see the annotation guide in the supplementary materials). In the pilot studies, small sets of data collected from Twitter were manually annotated using initial drafts of the annotation schema by two annotators (both with a linguistic background). Error analysis sessions were conducted in order to discuss and clarify unclear definitions and inconsistent judgments between annotators. The schema was then refined based on these discussions.

Two additional annotators were added to annotate more data, which we report on in Section 4.4. All four annotators are linguists familiar with WMNs. As in the pilot studies, an error analysis was conducted, which we discuss in Section 4.5).

Annotators were shown text of the tweets, one thread at a time, in the BRAT annotation tool (Stenetorp et al., 2012). Tweets were separated by a header that included the time of the tweet and the username of the tweet author. We displayed a maximum of 10 context tweets on either side of the candidate indicator.

Annotators were instructed to read the Twitter threads and select and classify text spans as different components of a WMN——as well as to determine whether or not an exchange as a whole was in fact a WMN. The four main points of interest, meant to be evaluated in order, during annotation were the WMN Type, Trigger spans, Anchors (Examples and Definitions), and instances of Grounding. While it was recommended that annotators examine these four points in order, we noted that it is completely acceptable and sometimes necessary to go back and forth to gain a better understanding of the thread.

**WMN Type** The search phrase (e.g., *what do you mean by*) was automatically pre-labeled as an Indicator to help the annotator find the intended focus of the example. Annotators were instructed to tag the Indicator span with the WMN Type of the dialogue as a whole. WMN Type consists of two decision points: First, the annotator must decide whether the thread is a WMN or not. If it *is* a WMN, it must then be classified as a non-understanding (NON) or disagreement (DIS).

**Trigger** The second task is to identify the word or phrase in question as the Trigger. Annotators must also label every other instance of the Trigger in the discussion, including anaphoric references. It is not necessary to link Triggers together with co-reference relations since it is implied.

**Anchors** The next step is to find the Trigger's Anchors and to distinguish between an Anchor's two types, Examples and Definitions. Relations are annotated with a link between the anchor and the Trigger o another Anchor, and the are marked with the polarity of the relation. An Anchor can also appear multiple times within a WMN, including anaphoric reference. In this case, these anchors are linked together using the co-reference relation. Annotators are instructed to try and leave negations out of the anchor and instead annotate the relationship as having negative polarity. When linking anchors, it is important which instance of the Anchor the link originates from, since this indicates which speaker is making the commitment and when. It is recommended that annotators use their best guess when identifying whether or not a URL (which could be an image or external link) is an Anchor and if so, its type based on the textual context.

**Grounding** Spans of text that explicitly state the speaker does (or does not) understand or agree with the previously offered example or definition must be annotated as Grounding. This span must be linked to the Anchor it refers to. The polarity link of a Grounding statement can be either positive or negative and between an Anchor and a Trigger or between two Anchors. In a non-understanding WMN, a grounding statement with a positive link indicates that the speaker understands the proposed relationship between the Anchor and the Trigger (or another Anchor). A negative link indicates that the speaker does not (or may not) understand the proposed relationship between the Anchor and the Trigger. In a disagreement WMN, a positive link indicates that the speaker agrees with or has adopted the proposed relationship between the Anchor and the Trigger. With a negative link, the grounding statement indicates that the speaker does not agree with or has not adopted the proposed relationship between the Anchor and Trigger.

### 4.3 Post-processing annotations

There are some discrepancies between the annotation schema and the WMN formalisation described in Section 3, mainly due to the fact the formalisation is comprised of abstract semantic units, while

the annotation is performed directly on the surface form of the WMN.

Text spans annotated as an *Anchor* (Example or Definition) were divided into equivalence classes, based on the co-reference annotations, which constitute the set of anchors in the formalisation. Spans annotated as Trigger were assumed to co-refer and the set of Triggers also constitutes an Anchor.

*Relation type* (Exa or Def in the formalisation) is coded as property of anchors in the annotation schema. In the pilot studies, we found that it was easier to decide the relational role of the anchor span before determining the polarity and target anchor. It is also more visually legible to separate the relation type (indicated by the color of the anchor span) and polarity (indicated by the color of the relation arrow). In theory, it would be possible for an anchor have multiple relational roles (imagine, for example, a WMN in which *insect* is used as both a partial definition of a *locust* and as an example of an *invertebrate*), but in practice this seems to be vanishingly rare (we have never observed it).

### 4.4 Results

In this section we report the results of the annotation study, particularly inter-annotator agreement.

We measured annotator agreement at two levels of description: the surface-form annotation, and then on the formal WMN representation extracted from the annotations. For agreement statistics, we report the proportion of agreed-upon items ($A_0$), as well as Cohen's kappa ($\kappa$) and Scott's pi ($\pi$).[6] Cohen's kappa computes expected agreement (the denominator) using annotator-level priors for the label distribution, whereas Scott's pi assumes a uniform distribution across annotators. Significantly higher $\kappa$ compared to $\pi$ would suggest that annotators have different priors for the category labels (Artstein and Poesio, 2008), but we don't observe that to be the case in any of the agreement statistics we measured.

First, we measured agreement on the dialogue level, namely, the classification of whether or not the dialogue was a WMN and if so, what type. Agreement was above chance, but (Table 1) with a substantial amount of disagreement. We discuss potential sources of disagreement in Section 5.

We measured agreement on span type at the token level. Tokenisation was performed post-hoc—annotators selected spans from the raw character-level text—but we consider a token to be part of a span if a majority of characters in the token overlap with it. This eliminates any artificial disagreements caused by, for example, missing the final letter in a word when selecting a span. We also consider it to be more representative than character-level agreement, which would be biased by longer words.[7]

| | $A_0$ | $\pi$ | $\kappa$ |
|---|---|---|---|
| WMN/Not | 0.71 | 0.40 | 0.40 |
| NON/DIN | 0.79 | 0.47 | 0.48 |

Table 1: WMN type agreement. *WMN/Not* measures agreement on whether or not the dialogue was a WMN, while *NON/DIN* (restricted dialogues both annotators agreed were WMNs) measures agreement on whether the WMN resulted from *non-understanding* or *disagreement*.

We found a moderate level of agreement on all span types except *grounding* (Table 2). Error analysis suggests that this may be primarily due to how much of a tweet the annotator considered to be a part of the grounding span. Additional guidance on this point in the annotation guide may help to raise the level of agreement.

| | $A_0$ | $\pi$ | $\kappa$ |
|---|---|---|---|
| Anchor | 0.93 | 0.59 | 0.60 |
| Trigger | 0.98 | 0.63 | 0.63 |
| Grounding | 0.98 | 0.22 | 0.22 |
| Overall | 0.87 | 0.64 | 0.64 |

Table 2: Token-level span type agreement. *Anchor* (both Definition and Example are considered *Anchor* here), *Trigger*, and *Grounding* only consider the binary choice of whether or not a token is of that type. *Overall* considers all three possibilities together.

At the level of the formal WMN representation, we are interested in whether annotators agree on whether and what kind of relations between anchors participants commit to at each turn, and when they explicitly indicate grounding of those relations. Computing agreement for relations and grounding requires that we align the anchors identified by the two annotators. For this, we take the bijection that maximizes token-level overlap of the spans associated with the anchors. This anchor mapping aligned an average of 89.1% ($\sigma$=19.2%) of anchors per dialogue (that is, on average 10.9% of anchors

[6] $A_0$ is the numerator for both $\kappa$ and $\pi$.

[7] We used the NLTK (v.3.6.2) regex-based TweetTokenizer.

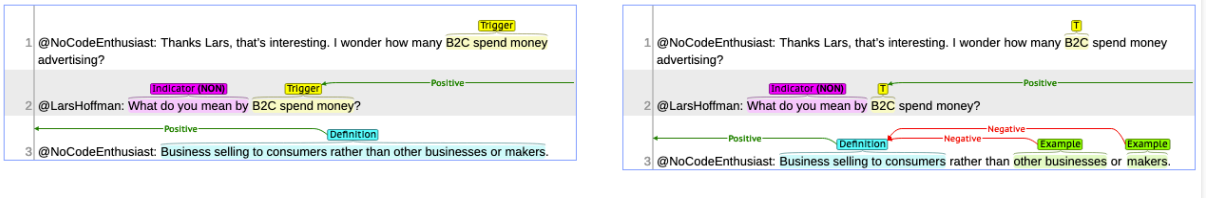


Figure 1: Annotations from two annotators, showing disagreement in the extent of the trigger phrase and anchor structure.

had no counterpart in the other annotation).

For relations, we considered each *potential* relation at each turn; that is, for turn $i$, we consider each pair of anchors (including the trigger), $\{(a, b) \in \boldsymbol{A}_i \times \boldsymbol{A}_i \mid a \neq b\}$ (with the caveat that $\boldsymbol{A}_i$ only includes *aligned* anchors, since there is no possibility for agreement on unaligned anchors). Annotators agree if they both created a relation (with the same relation type and polarity) from an $a$-span originating in turn $i$ to an $b$-span (regardless of where)—or if they both created no relation at all for that pair. As with the token-level statistics, $A_0$ is quite high, since relations are sparse, relative to all the opportunities for a relation to be created, but the chance-adjusted scores are also reasonably high (Table 3).

For grounding, for each turn $i$ (starting with $i = 1$), we considered each aligned anchor that both annotators agreed was mentioned in turn $i - 1$. Annotators agree if they both thought that the current speaker grounded (with the same polarity) a relation originating in that anchor—or if they both thought no such grounding occurred. Agreement is lower than for anchor relations, but still well above chance (Table 3).

| | $A_0$ | $\pi$ | $\kappa$ |
|---|---|---|---|
| Relation | 0.93 | 0.69 | 0.69 |
| Grounding | 0.88 | 0.58 | 0.59 |

Table 3: Turn-level agreement on relation type and grounding polarity for possible relations and grounding.

### 4.5 Error analysis

After annotating the examples, we conducted some post-hoc discussions in which the annotators attempt to ascertain the reason for certain discrepancies. Based on these discussions, we make suggestions for improvements to the annotation protocol, which should aid in future efforts to annotate WMN. Further observations about the phenomenon of WMN, which came to light in these conversations, can be found in Section 5.

**WMN Type** The phrase *what do you mean by* is often used in a rhetorical way (i.e., not as a genuine question or clarification request), but it can be difficult to determine whether the speaker's objection to using a word to describe some situation under discussion is a disagreement about the meaning of the word (DIN) or a disagreement about the nature of the situation under discussion (not a WMN). The decision could be clarified by emphasizing the *results* of the indicator phrase: Does the meaning of the word subsequently become at-issue? When non-understanding or disagreement is indicated but no meaning negotiation results, this is typcially not considered a WMN (Varonis and Gass, 1985; Myrendal, 2015), but giving such "declined WMNs" their own category could result in better agreement.

**Anchor spans** Analysis revealed two kinds of discrepancy in anchor spans: (1) where the annotators disagreed on whether something was an anchor, or how much of the text referred to the anchor (reflected in token-level agreement, Table 2), and (2) where the annotators disagreed on whether something was one anchors or two (reflected mainly in the failure to find a bijection between the two annotated sets of anchors).

A particularly notable discrepancy of the first kind involves the extent of the trigger phrase, since the speaker will sometimes repeat some context around the trigger to help locate it in the previous utterance. This can raise the question of how much of what they repeated is context and how much is the trigger. One strategy for annotators could be to observe what is *actually negotiated* subsequent to the indicator, although this too can be ambiguous.

Another common discrepancy was that one annotator would annotate multiple anchors, where another would find only one (see Figure:1).

**Relation types** While agreement on relation type (annotated as anchor span type) was fairly good, there were a few cases where adding more relation types could improve clarity. *Contrasting* is a common pattern in WMNs where the trigger word is compared to an alternative that the speaker thinks better describes the situation under discussion (Myrendal, 2019): *x is really more of a Y than a Z*. In the annotation guide, we suggested such examples be annotated with two relations: $\mathsf{Exa}^{+}(Y, x), \mathsf{Exa}^{-}(Z, x)$, but it could also be its own ternary relation that is interpreted using $\delta$ and $\epsilon$, as in Larsson and Myrendal (2017).

## 5 Discussion and conclusion

We conclude by offering some observations on the WMNs in our Twitter corpus, and discussion on the implications these observations may have for negotiated meaning more broadly.

**Speaker meaning/token meaning** As mentioned in Section 4.5, it was often unclear whether *what do you mean by X* was asking what the speaker understands *X* to mean in general, or what they were *using X* to mean in a particular context.[8] This is perhaps related to the phenomenon where the indicator repeats a whole sentence, but the negotiation focuses on one word or short phrase: Since questions about sentence meaning are necessarily about speaker meaning, including the sentence in the indicator may clarify that the question is about speaker meaning. Clark (1996)'s hierarchical grounding schema, makes the distinction between grounding on the level of *signal meaning* and grounding on the level *uptake* (speaker meaning or illocutionary act). When a WMN is focused on resolving a non-understanding (NON), the issue can be either with the signal meaning or with uptake, however a disagreement (DIN) about how a word is used is necessarily a disagreement about its *meaning potential* (Linell, 2009)—it doesn't make sense to disagree *that* someone meant something, only *how* they went about meaning it.

**Social and cultural context** Many of the WMNs in our corpus involved politically or socially controversial topics and the moves made by the participants often required some understanding of the social context in which the conversation was taking place. Consider the example in figure 2: Interpreting *convicted by the court* as providing a negative example of a *delusional conspiracy*, requires understanding the role of *conspiracy* in Indian political discourse, what *Congress* refers to (a political party) and even the political alignment implied by *Saffron* in one of the usernames.

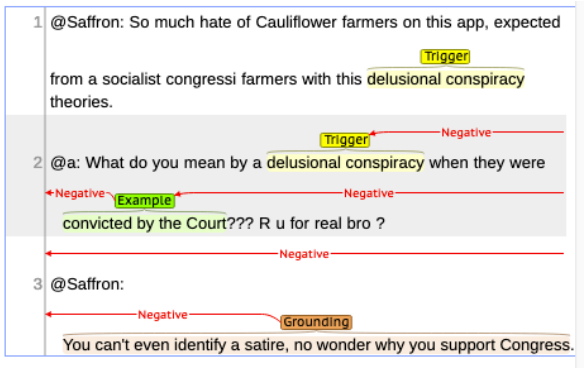


Figure 2: Post-hoc annotation provided by an annotator familiar with Indian social media political discourse. The original annotators of this example, lacking the background knowledge, had different interpretations.

**Agreement and reliability** As the cultural context example demonstrates, annotator disagreement doesn't *necessarily* imply that the annotation schema is incorrect or doesn't reflect the underlying phenomenon. In that case, one of the annotators lacked the context to interpret the WMN correctly, but it is possible for WMNs to be ambiguous (open to multiple possible interpretations), even when both have sufficient background knowledge. Reflecting these different interpretations can make this formalisation a useful tool for analysis, just as first-order logic is a useful tool for analysing certain classes of ambiguous sentences.

Taking that for granted, and considering our somewhat mediocre annotator agreement scores, what can we conclude about this formalisation and annotation schema? Is it in some sense *correct*? The only way to know is probably to continue using it (and where possible, improve upon it)—to carry out further annotation studies on conversational data from different sources, formulate and test hypotheses, and eventually attempt to train artificial agents capable of WMN.

As explicit meta-linguistic discussions, WMNs have potential as window into the processes of semantic alignment, acquisition, and change more generally. By modeling WMNs, we hope to develop conceptual frameworks that apply to the dynamics of lexical semantic resources more broadly.

[8] See also: Myrendal (2019) *general* versus *specific explicifications*.

## References


Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher and Alex Lascarides. 2008. Commitments, Beliefs and Intentions in Dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Eve V. Clark. 2007. Young Children's Uptake of New Words in Conversation. *Language in Society*, 36(2):157–182.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. 2014. Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue Acts and Updates for Semantic Coordination. In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 52–59. ISCA.

Per Linell. 2009. *Rethinking Language, Mind and World Dialogically : Interactional and Contextual Theories of Human Sense-Making*. Information Age Publishing.

Michael H. Long. 1996. The Role of the Linguistic Environment in Second Language Acquisition. In William C. Ritchie and Tej K. Bhatia, editors, *Handbook of Second Language Acquisition*, pages 413–468. Academic Press, San Diego.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*, first edition edition. Oxford University Press, Oxford, United Kingdom.

Shiwali Mohan, Aaron Mininger, James Kirk, and John E. Laird. 2012. Learning Grounded Language through Situated Interactive Instruction. In *2012 AAAI Fall Symposium Series*.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. PhD Thesis, University of Gothenburg, University of Gothenburg.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication - Jenny Myrendal, 2019. *Discourse Studies*, 21(3):317–339.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

E. M. Varonis and S. Gass. 1985. Non-native/Non-native Conversations: A Model for Negotiation of Meaning. *Applied Linguistics*, 6(1):71–90.