Classification of low-agreement Pronouns through Collaborative Dialogue: A Proof of Concept

Janosch Haber and **Massimo Poesio**
Queen Mary University of London
{j.haber|m.poesio}@qmul.ac.uk

Abstract

When evaluating model performance on automated annotation tasks such as anaphora resolution and specifically pronoun resolution, the gold standards often postulate a single correct referent for each referring expression. Previous research on annotator disagreement however found that in some cases there might not actually be a single correct referent, but rather multiple acceptable alternatives - or no specific solution at all. In this paper we aim to extend the study on pronouns with low annotator agreement by investigating how speakers react to various occurrences of *it* pronouns in a collaborative dialogue setting. We propose that different types of ambiguity and under-specification can be distinguished based on whether or not speakers discuss the resolution of a given pronoun during the task, and whether or not they agree on a referent for it afterwards. Applied to corpus samples which previously received low annotator agreement, we propose that this methodology provides a comparatively cheap means to aid distinguishing annotation noise from genuine classification difficulty.

1 Introduction

For decades now it has been practice to train or at least evaluate Natural Language Processing (NLP) systems on annotated datasets. Linguistic annotations are traditionally compiled using either a number of domain experts or trained annotators, or, if annotations are required on a larger scale, in recent years are increasingly crowd-sourced from layman judgements through Games-with-a-purpose (GWAPs, see von Ahn et al., 2006; von Ahn and Dabbish, 2008), Citizen Science Projects and other paid and unpaid online applications. As especially the latter approach tends to lead to noisy data, it is customary to collect multiple labels for every item and establish the final gold label by aggregating or adjudicating the various collected annotations. This means that the resulting gold standard - based on which models are trained and against which systems will be tested - postulates a single correct label for each item only. Focusing on anaphora resolution, the task of resolving the dependencies of expressions that rely on some form of linguistic context (Poesio et al., 2016, p. 24), empiric research on annotator disagreement however revealed that not in all cases is data collection noise to blame for obtaining a range of alternative labels, but rather that sometimes the different labels are due to genuine disagreement among the annotators (see e.g. Poesio and Artstein, 2005a; Versley, 2008).

1.1 Causes of Annotator Disagreement

Genuine annotator disagreement on pronoun resolution can be the result of a number of factors that play a role in the processing of a anaphoric expressions. Here, we will focus on three types of pronouns that have been found to make up a substantial part of low-agreement samples: ambiguous, under-specified and expletive pronouns. The term *ambiguity* is often used to to indicate expressions with unclear meaning, but in a more strict sense only refers to expressions whose interpretation can be selected from a discrete set of alternatives, separating the phenomenon from instances of vagueness (Pinkal, 1985; Poesio, Forthcoming). If an anaphoric expression is ambiguous in the narrow sense, we expect annotations to resemble a distribution of labels according to the different interpretations' dominance, and the collected labels should therefore be spread over a few alternative candidates. *Under-specification* is used to classify instances of dialogue or written text for which a listener or reader does not seem to build a complete interpretation (Poesio and Reyle, 2001). First evidence of under-specified expressions had been

reported in Poesio et al. (2006) who observed that participants in a collaborative dialogue task used pronouns *it* and *that* to refer to complex objects that were not explicitly introduced to the discourse before. While often syntactically infelicitous or semantically ambiguous, under-specified expressions have usually been found not to interfere with task success, originating the notion of *justified sloppiness* to account for infelicitous expressions that do not negatively affect the interaction and in some cases might even increase task efficiency (Piantadosi et al., 2012). If the referent of a pronoun is under-specified in this way, we expect that a traditional annotation setup will yield a wide spread of candidate referents as annotators are likely to select different combinations of available anchors to proxy the under-specified referent not explicitly available in the linguistic context. *Expletive*, or *non-referring* pronouns lastly do not have a linguistic referent at all, but rather function as dummy pronouns catering to grammatical requirements of syntactic composition. If identified correctly, expletive pronouns should not be assigned a referent, but annotation noise may lead to a small set of spuriously linked anchors.

1.2 Discussing Disagreements

To further investigate the notion of genuine annotator disagreement caused by ambiguity, under-specification and expletives as opposed to annotation noise, we propose to extend the analysis of items obtaining low annotator agreement in crowd-sourced corpora like Phrase Detectives (Poesio et al., 2019) by testing them in a dedicated dialogue task. The reasoning behind this proposal is as follows: In crowd-sourced annotation tasks, annotators usually are not aware of their disagreement as it emerges only as a result of their individual label assignments. This means that, for example, disagreements on difficult items can arise because i) annotators are not aware of the ambiguity or under-specification of a given expression and label it in different ways, or ii) they might be unable to express ambiguity or under-specification in a canonical way corresponding to those of other annotators. If annotators however get an opportunity to discuss items, they could be made aware of alternative readings, either because those are directly proposed by other annotators, or because of disagreements in interpretation when comparing individually assigned labels, and referents can be assigned collaboratively. Observing these referent discussions in the dialogue task thus could be used as an indicator of item difficulty and samples causing them marked accordingly, whereas low-agreement items not discussed in this setting are likely to be a result of annotation noise.

Explicit discussion of annotation labels is usually used in the adjudication of expert annotators' labels (see for example Finlayson and Erjavec, 2017), but here we propose to apply this paradigm implicitly in a fully crowd-sourcing setting by having crowd workers discuss text segments containing pronouns that received low agreements on their crowd-sourced referent labels. To do so, we developed a novel two-staged dialogue task. In the first part of the task, a short text passage containing a low-agreement pronoun is split into segments, which are presented to a participant pair in random order. The participants are then asked to collaboratively re-produce the original passage by re-ordering the segments. This requires the participants to establish the referents for at least those pronouns central to the understanding of the passage, and is likely to prompt a discussion if participants are in doubt whether they share the same interpretation of these pronouns. Once the participants agree on an ordering of the text snippets, they enter the second part of the task where - now individually - they are asked to assign referents to the pronouns occurring in the text. We propose that combining observations on whether or not participants discussed a pronoun in the first part and whether or not they agree on the individually assigned referents in the second part will yield insightful information for a classification of low-agreement items:

1. If low agreement was caused by annotation noise in the original data collection, we expect participants in our setup to agree on the referent in the second stage, but not necessarily discuss the pronoun during the initial collaborative re-ordering phase, because it likely that the pronoun is actually straightforward to resolve and doesn't require deliberation.

2. If low agreement is due to referent ambiguity, we expect participants to discuss the pronoun and its resolution during the first part of the task, and determine how they will interpret it for the re-ordering of the passage. This means

that they should agree on the assigned referent in the second part of the task.

3. If an under-specified item caused low annotator agreement, we expect that participants in the dialogue task will not extensively discuss the pronoun during the first stage, as the *justified sloppiness* hypothesis predicts that the pronoun's referent does not need to be fully specified to complete the interaction objective (here re-ordering the text segment) - but we also expect that participants are unlikely to agree on a referent in the second part of the task, as they never made a referent explicit.

4. Lastly, non-referring expletive *it*s are expected to produce high agreement (on not assigning a referent) during the second part of the task, but to be unlikely to trigger any discussion of referents during the first part.

In this paper we present a pilot study intended as a proof of concept for the classification of low-agreement items through collaborative dialogue. We conducted a small-scale pilot run in the lab rather than applying actual crowd-sourcing techniques, and used a longer input text that specifically contains instances of all three types of problematic *it* pronouns in order to quickly collect dedicated data for all cases. We show that the collaborative task allows us to classify the three types of *it* pronouns occurring in the text using the output of the two stages of the task, validating the task setup, and outline the further development of the task into a proper crowd-sourcing tool.

2 Related Work

The dialogue task presented in this paper draws from previous work surrounding the discovery of under-specified expressions, especially observed in dialogue settings, and evidence collected to support the idea that annotator disagreement is not necessarily data collection noise, but rather can sometimes be used as valuable data for improving computational attempts to anaphoric resolution.

2.1 Under-specification

Investigating the transcripts of the TRAINS corpus (Gross et al., 1993; Allen and Heeman, 1995), a set of dialogues collected in an wizard-of-oz setup for planning alongside an automated system, Poesio et al. (2006) observed that in some cases participants - as well as the instructor posing as the automated planning system - used anaphoric expressions that seemed to not have a clear referent in the preceding discourse. Consider for example excerpt (1) where participant A is tasked to send a boxcar to a train station in Corning. B is the instructor posing as an automated planning system:

(1) A: Can we kindly hook up, uh, engine E2 to the boxcar at Elmira and send it to Corning as soon as possible please?
B: Okay.
A: Do let me know when it gets there.
B: Okay, it should get there at 2AM.

Resolving the highlighted *it* pronouns traditionally would require the selection of one of the two available anchors *boxcar* or *engine*, rendering them ambiguous in this context. Speaker B however does not seem to raise a complaint over this ambiguity but continues the interaction as if the referent was identified without any issues. Poesio and Reyle (2001) propose a number of alternative ways to explain speaker B's processing of the under-specified reference, but independent of how exactly this processing occurs, the interaction can be presumed successful, as the plan formulated by speaker A is understood and correctly executed by speaker B. This led the authors to introduce the concept of *justified sloppiness*, arguing that expressions can be left under-specified when their realisation is irrelevant to the conversation goal. Versley (2008) later adopted this notion of under-specification and formulated the *Generalised Sloppiness Hypothesis* based on Asher and Pustejovsky (2006)'s dot-object formulation for complex types. According to this model, anchor *engine E2* is represented as an under-specified complex object with the two combined aspects *engine•train*, and *boxcar* as *boxcar•train*, respectively. The instruction *Send ? to Corning* then selects for the *train* aspect of both anchors only, allowing them to be combined into a singular mereologic structure that can be referred to with *it*.

Taking a different approach, Recasens et al. (2011) propose that co-reference does not require identity of reference, but rather that items which are merely "near-identical" can be perceived as co-referent. This hypothesis builds on previous work by Swets et al. (2008) who observed lingering effects of syntactic garden-path sentences and

proposed that the interpretation of those structures in online processing is not fully disambiguated but rather "good enough" to continue reading.

2.2 Annotator Disagreement as Data

Concerning the use of annotator disagreement to enrich model input istead of disregarding it as data collection noise, Plank et al. (2014b) for example observed that annotator disagreements in part-of-speech (POS) tagging were systematic across domains and to a certain extend also across languages, and that a majority of high-disagreement items actually stemmed from linguistically debatable instances. Utilising expected annotator disagreement for hard cases as a weight in the loss function of an automatic POS tagger, Plank et al. (2014a) subsequently recorded gains in model accuracy and downstream task performance. Others, too, have suggested to specifically identify and adapt the treatment of ambiguous and subjective cases in corpora from a range of domains and tasks, including for example image segmentation (Firman et al., 2018), judgements about visual scenes (Sharmanska et al., 2016), dialogue addressing annotations (Reidsma and op den Akker, 2008) or classifying expressions as semantically old or new (Klebanov and Beigman, 2014).

Battling the noise in crowd-sourced annotations, a growing body of recent work presents approaches to estimate annotator reliability and use it as a metric to weigh annotations when aggregating gold labels from annotations, as for example in Aroyo and Welty (2013, 2015); Schüller (2018) and Paun et al. (2018).

3 Method

The pilot study presented in this paper was developed to establish a proof of concept before applying the dialogue task to a large-scale crowd-sourced validation of previously collected annotations. This means we tested the task with a small number of participants in a lab setting rather than online, and instead of using actual text segments containing low agreement items from corpora such as ARRAU (Poesio and Artstein, 2005b) or Phrase Detectives (Poesio et al., 2019), we created a longer sample text that contains instances of all three of the *it* pronoun types mentioned earlier.

The pilot text segment takes the form of an interview with a fictional band member. The interviewee reports about a failure in guitarist *Sophie*'s equipment during a live show and how she and the sound engineer *Megan* attempt to fix it. The story predominantly uses *she* pronouns to refer to either of these two characters, and reference needs to be established in order to determine the actors for most part of the story. In context, the *she* pronouns are unambiguous, but due to the initially random ordering of snippets in the task, they remain ambiguous until placed in order. Besides these personal pronouns, the story includes *it* references that in the same way are initially ambiguous between a number of objects but can largely be resolved once context is established, non-referring *it* pronouns (which remain expletive in established context), and *it* references to mereologic constructions that are not explicitly mentioned in the text. The text segment is divided into an introduction paragraph, ten snippets containing mostly a single sentence each, and a short ending paragraph.[1] The ten snippets are indicated with a random letter to aid the transcription of deictic references and show pronouns underlined and numbered with a small sub-script.

3.1 Task Setup

Two participants with university-level, self-reported native English language skills were instructed that they were to participate in a two-staged task. The first part would require them to collaboratively re-construct a story which was cut into a number of snippets, and the second part to answer some questions about the re-ordered story. The participants were then seated next to one another at a table in a small conference room. Paper prints of text snippets containing the start and ending paragraph of the story were laid in the middle of the table, as well as a simple diagram showing the story's setting with colour-coded circles for characters and icons for objects. Each participant was then given five random snippets. The participants were asked to first read the story's start and their own snippets and then collaboratively order the ten snippets below the story start. Participant interaction was audio recorded during the task and the instructor took notes of deictic references to task elements to aid the transcription of the interaction.

Once the participants signalled that they agreed on a snippet ordering, both were handed mention sheets listing all of the text's underlined pronouns. The participants were then asked to individually

[1] See Appendix A.1 for the full pilot text.

assign referents to the pronouns by either writing down the referent or assigning the same colour as indicated in the diagram. This second stage was completed without participant interaction. After participants finished filling in the mention sheets, the instructor compared the sheets and asked questions about some of the pronouns, especially if they were not assigned a referent or if the participants disagreed on referents, in order to make sure that the mention sheets represented the resolution of pronouns as intended by the participants. The mention sheets were then compared for differences in the resolution of pronouns to establish agreement and disagreement on referents by the end of the interaction.

3.2 Annotation Scheme

Besides the participants' agreements and disagreements on the particular referents of the pronouns we are interested in whether or not a pronoun's referent was discussed during the interaction in the first stage. Since this is less straightforward to assess than participant agreement, we instructed trained annotators to go through the transcripts and mark all instances of pronouns being discussed based on a simplified account of core speech acts *propose*, *discuss* and *accept* (e.g. Poesio and Traum, 1997) combined into two different cases:

1. **Pronoun resolved.** A pronoun is considered to be resolved during the interaction if i) its referent was either explicitly discussed by the participants or ii) a referent was indirectly established by snippet ordering and accepted by both participants.

   A referent is considered to be discussed explicitly if either i) a participant refers to the pronoun and then proposes a referent for it, ii) a participant re-formulates the snippet to contain a referent instead of the pronoun originally used in the snippet, or iii) a participant refers to a snippet that contains a pronoun and proposes that it "is" or "is done by" or "is about" a certain referent, and the referent is accepted by the other participant. A referent is established indirectly if either i) a participant reads a block of snippets containing a given pronoun, indicating that they form a coherent unit, and the pronoun is ambiguously resolved in the newly established context, or ii) a participant proposes that a snippet should precede or succeed another snippet or block of snippets which form a context that unambiguously resolves the pronoun, and the proposal is accepted by the other participant. A proposed referent finally is considered as accepted if the other participant either i) explicitly agrees with the proposal or ii) changes the topic of the conversation, indicating that the proposal does not require further discussion.

2. **Pronoun unresolved** A pronoun is considered to remain unresolved after the interaction if a proposed referent was rejected by the other participant or the pronoun and its referent were not discussed during the interaction.

The annotators were not made aware of the proposed distinction between ambiguous, non-referring and under-specified *it* pronouns.

4 Results

During the pilot run we collected and transcribed the conversations of 11 participant pairs discussing the re-ordering of the pilot story. Conversations took between 5 and 20 minutes (mean = 11:04 min, std = 5:58 min), with transcripts containing between 38 and 151 utterances (mean = 109.09, std = 42.49). The mention sheets individually filled in by the participants after their interaction show an overall agreement on 60.80 % for the referents of the pronouns used in the story's snippets, with agreement rates for the different pairs of participants ranging between 31.25% and 87.50%. Noticing these high fluctuations, we decided to exclude transcripts from further analysis if participants did not agree on the referents of at least 7 of the 8 personal pronouns in the story. Since the personal pronouns need to be resolved collaboratively during the interaction in order to establish a common ground about the events in the story, we argue that in interactions where participants fail to agree on the referents of these mentions, it is unlikely that they both correctly understood the task or the story. Only 7 out of the 11 participant pairs met this requirement, resulting in a filtered dataset with conversations lasting between 6 and 18 minutes (mean = 10:14 min, std = 5:08 min), transcripts containing between 64 and 151 utterances (mean = 108.29, std = 37.51) and an overall agreement on referents in 71.43% of the cases, with a lowest individual score of exactly 50%.

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	100.00%	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%	98.21%
Ambiguous *it*	100.00%	0.00%	100.00%	33.33%	0.00%	66.67%	100.00%	57.14%
Non-referring *it*	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	57.14%
Under-specified *it*	33.33%	0.00%	33.33%	0.00%	33.33%	0.00%	66.67%	23.81%
Total	87.50%	50.00%	81.25%	68.75%	56.25%	75.00%	81.25%	71.43%

Table 1: Referent agreements per participant pair and pronoun type in the data collection pilot run.

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	84.38%	87.50%	87.50%	84.38%	78.13%	71.88%	75.00%	81.25%
Ambiguous *it*	100.00%	91.67%	91.67%	83.33%	75.00%	75.00%	75.00%	84.52%
Non-referring *it*	87.50%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	98.21%
Under-specified *it*	91.67%	100.00%	91.67%	91.67%	66.67%	58.33%	91.67%	84.52%
Total	89.06%	92.19%	90.63%	87.50%	78.13%	73.44%	81.25%	84.60%

Table 2: Annotator agreement on their assessment whether or not a participant pair resolved a given pronoun during their interaction. Results grouped per transcript and pronoun type.

4.1 Evaluation of Agreements

A detailed overview over the agreements of the different participant pairs and pronoun types is shown in Table 1. In the filtered dataset, all but one participant pair agree on the referents of all eight personal pronouns in the story, with the remaining pair disagreeing only on one item, resulting in an average agreement rate of 98.21% for the personal pronouns. This is significantly higher than for any of the *it* pronoun types (z-score for population proportions p-value < 0.01). The total agreement on both ambiguous and non-referring *it* pronouns in turn is 57.14%, significantly higher than the total agreement rate of 23.81% for the under-specified *it* pronouns (p-values = 0.028 and 0.046, respectively). Note that agreement is regarded independent of whether participants assigned the intended referent, and agreement on non-referring pronouns was assumed if neither participant assigned a referent as well.

4.2 Evaluation of Annotations

We collected annotations from four trained annotators instructed to follow the annotation guidelines as stipulated above. The average observed agreement concerning the resolution of pronouns by participants reached 0.73, with inter-annotator agreement however only scoring a Cohen's κ (Cohen, 1960) of 0.47, and pairwise κ scores ranging between 0.31 and 0.75. Table 2 shows in more detail the observed annotator agreement concerning whether or not a pronoun's referent was resolved by the participants in the interaction, differentiated by participant pair and pronoun type. Annotator agreement varies between different participant pairs, ranging from 73.44% to 92.19%, but is almost identical among the different pronoun types, with only non-referring pronouns showing a significantly higher annotator agreement rate than any of the other types.

4.3 Evaluation of Referent Discussion

In order to assess whether a pronoun's referent should be considered as resolved during the interaction, we aggregated annotator labels through majority vote. Considering the low annotator agreement observed above, we did so for all four annotator labels and the labels obtained from the two annotators with highest agreement only and compared results. This filtering however only had a minimal effect, and consequently it was decided to use the full set of annotations and keep its expressiveness. Table 3 shows the ratios of resolved pronouns per transcript and pronoun type based on the aggregated labels of all four annotators. According to the annotators' judgement, ambiguous *it* pronouns are collaboratively resolved most often during the interactive part of the dialogue task, although not significantly more often than the ambiguous personal pronouns (z-score p-value = 0.1436). Non-referring *it* are annotated to have never been collaboratively resolved during the interaction, a finding that directly corresponds to the nature of the pronoun type. Under-specified pronouns lastly are discussed significantly less than the personal and ambiguous *it* pronouns (p-values < 0.01), but not significantly more often than the non-referring pronouns (p-value = 0.1391). Note that no observed

Pronoun Type	1	2	3	4	5	6	7	Total
Personal Pronoun	100.00%	100.00%	75.00%	75.00%	75.00%	87.50%	62.50%	82.14%
Ambiguous *it*	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	100.00%	95.24%
Non-referring *it*	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Under-specified *it*	66.67%	0.00%	33.33%	0.00%	0.00%	0.00%	0.00%	14.29%
Total	81.25%	68.75%	56.25%	56.25%	56.25%	62.50%	50.00%	61.61%

Table 3: Pronoun referent resolution per participant pair and pronoun type as judged by independent external annotators.

discussion of a referent does not entail that participants did not assign a referent at all. This becomes especially clear when keeping in mind that referents for ambiguous and most under-specified pronouns actually were assigned by the individual participants during the second part of the task.

4.4 Combination of Factors for Classification

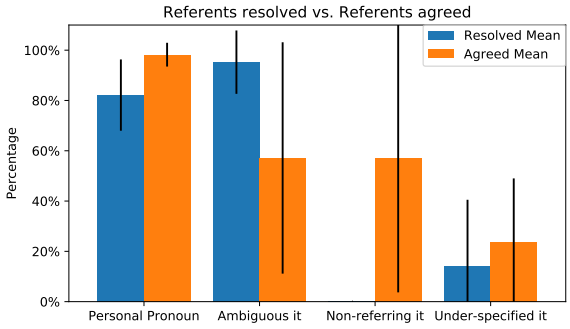

Figure 1: Distribution of pronouns resolved during the interaction (blue) and referents agreed afterwards (orange), grouped per pronoun type.

Figure 1 shows the overall distribution of pronouns collaboratively resolved by participants during their interaction in the first part of the task (blue) compared to the ratio of referent agreement in the second part (orange). The figure suggests that there are structural differences in participants' reactions to the four different types of pronouns tested, with (ambiguous) *she* pronouns showing almost perfect agreement with a relatively high amount of discussions, ambiguous *it* pronouns showing more discussion but less agreement, non-referring *it* pronouns exhibiting mediocre agreement but no discussion whatsoever, and under-specified *it* pronouns displaying low levels of both, discussions and agreement. When tested with χ^2, the resulting test statistic is improbably large compared to the χ^2 distribution given six degrees of freedom ($\chi^2 = 48.23$, p-value < 0.01), indicating that the distribution of the four combinations of metrics significantly differ between the three pronoun types. Combining information on whether pronouns were discussed in the first part and whether they were agreed upon in the second part thus seems to uniquely identify the three different types of pronouns, as supported by the contingency table in Table 4.

Pronoun Type	¬R ¬A	R ¬A	¬R A	R A
Ambiguous *it*	0	9	1	11
Non-referring *it*	6	0	8	0
Under-specified *it*	14	2	4	1

Table 4: Contingency table containing counts of instances where pronoun referents were collaboratively resolved (R) during the first part of the task and agreed upon (A) in the second part of the task.

5 Discussion

While the obtained results seem promising prima facie, there are some factors to be taken into account when interpreting them with regard to a classification of low agreement items, and in perspective of transforming the pilot layout to a crowd-sourcing setting.

5.1 Pilot Shortcomings

Ambiguous pronouns. The ambiguous personal and *it* pronouns in the pilot text were not ambiguous in context but rather ambiguous as to their referent prior to putting them in order. When using previously annotated texts as input data, referents can remain (and in many of the genuine low-agreement items will remain) ambiguous after assigning a snippet order. We however argue that this does not have a major impact on the task setup, as the important factor is whether or not the pronoun and its referent were discussed during the first part of the task. Additionally, it is likely that participants will collaboratively decide on a specific interpretation even for genuinely ambiguous pronouns in order to ground their interpretation in the discussion of snippet ordering.

Snippet ordering. In the pilot setup, we analysed

transcripts independent of the fact whether participants re-created the intended ordering of the text segment or not. Considering however that the pilot text was created so that it could generate a number of sensible orderings, and that a future data collection using crowd-sourcing will have much shorter input texts (see section 5.2), we believe that this does not discount our proof of concept.

Annotation of implicit referent proposals. The annotation guidelines for marking the discussion and resolution of referents during the interaction allows for *implicit* proposals through snippet ordering, i.e. when a snippet containing a pronoun is proposed to succeed a snippet containing an anaphoric referent. We realised that these cases were difficult to spot even for trained annotators, especially considering that under-specified and non-referring pronouns do not actually refer to a preceding anchor. When applied in a crowd-sourced manner, implicit proposals therefore will need to be re-defined or disregarded entirely.

Rejection rate. A last issue we cannot readily explain is the high number of participant pairs who did not agree on referents for the *she* pronouns crucial for an understanding of the pilot text. It appears that in some cases the task of re-ordering snippets overshadowed the processing of the text itself, in some cases participants had questionable text comprehension skills, and in some cases the the task might have been misunderstood. We however expect that a simplified setup with shorter input texts will mitigate this effect.

5.2 Transformation to Crowd-Sourcing

Considering the pilot to provide proof of concept, we believe that the developed dialogue task could be applied in a crowd-sourcing setting to classify low-agreement items in previously annotated corpora. A number of modifications however needs to be made to transform the pilot setup into a data collection tool suitable for crowd-souring. Specifically, we need to automate the generation of input segments from previously annotated corpora, and transform to an online setting i) the collaborative dialogue phase, ii) the individual referent assignment phase, and iii) the annotation of whether or not a pronoun is discussed in the dialogue stage. Regarding the generation of input texts, further work is needed to investigate whether the contexts of low-agreement items derived from corpora like ARRAU or Phrase Detectives can be automatically reduced to a length feasible for the collaborative re-ordering task. Considering the conversion of the first stage of the task, we already started developing a first digital version of the pilot using crowd-sourcing tool SLURK (Schlangen et al., 2018). The tasks' user interface was re-designed to contain a private section holding the snippets and a shared section to collaboratively arrange them, as well as a chat box, moving from spoken to written dialogue to eliminate the need for transcription and making the task more accessible. We are currently working on updating the instructions to further gamify the task to increase worker commitment and internal motivation. We see few issues with implementing the individual second part of the task, but consider obtaining independent annotations concerning the discussion of referents through crowd-sourcing to be the most daunting part of converting the task setup. As we already observed relatively low annotator agreement even among linguistically trained annotators, the best strategy for simplifying the annotation task for crowd-sourcing remains a central open issue to be addressed by future work.

6 Conclusion

The data collected through the pilot setup presented in this paper indicate that three different types of pronouns that are likely to cause low annotator agreement in classic annotation tasks can be differentiated based on observations obtained from a dedicated dialogue task which implicitly requires participants to collaboratively resolve their referents. Ambiguous expressions are found to be discussed and resolved more often than other types of pronouns, with the referents of non-referring pronouns usually being agreed upon without explicit discussion and under-specified pronouns triggering relatively little discussion but also leading to low levels of agreement. In future work, this combination of observations could be used to augment current gold standard annotations by classifying low-agreement items as either caused by genuine item difficulty or annotation noise, and provide richer data for training and testing NLP systems. Our findings also provide some support for the hypotheses of *justified sloppiness* and *good enough representations*, suggesting that referents of under-specified pronouns do not need to be fully resolved in order to be interpreted similarly enough to not hinder understanding and ultimately task success.

Acknowledgements

The work presented in this paper was supported by the DALI project, ERC Grant 695662. The authors would like to thank Derya Çokal and Andrea Bruera for their input, and the anonymous reviewers for their feedback.

References

Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM*, 51(8):58–67.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A Game for Collecting Common-sense Facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 75–78, New York, NY, USA. ACM.

J Allen and P A Heeman. 1995. TRAINS Spoken Dialog Corpus [CD-ROM]. *Philadelphia, PA: Linguistic Data Consortium*.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Nicholas Asher and James Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1):38Asher, N., & Pustejovsky, J. (2006). A type comp.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mark A Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer.

Michael Firman, Neill DF Campbell, Lourdes Agapito, and Gabriel J Brostow. 2018. Diversenet: When one right answer is not enough. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5598–5607.

Derek Gross, James F. Allen, and David R. Traum. 1993. The trains 91 dialogues. Technical report, University of Rochester, USA.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396.

Silviu Paun, Bob Carpenter, J D Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280 – 291.

Manfred Pinkal. 1985. *Logik Und Lexikon: Die Semantik des Unbestimmten*. De Gruyter.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio. Forthcoming. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *In Proc. of the Corpus Linguistics Conference*.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Massimo Poesio and Uwe Reyle. 2001. Underspecification in anaphoric reference. *Fourth International Workshop on Computational Semantics (IWCS-4)*.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution*. Springer.

Massimo Poesio, Patrick Sturt, Ron Artstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*.

Massimo Poesio and David R Traum. 1997. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. slurk–a lightweight interaction server for dialogue experiments and data collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (Aix-Dial/semdial 2018)*.

Peter Schüller. 2018. Adjudication of coreference annotations via answer set optimisation. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(4):525–546.

Viktoriia Sharmanska, Daniel Hernández-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2202.

Benjamin Swets, Timothy Desmet, Charles Clifton, and Fernanda Ferreira. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3-4):333–353.

A Appendices

A.1 Pilot Text Segment

The text segment used in the pilot study contained the following snippets (presented in intended order):

Introduction
MusicManiac Magazine interviewed art-rock trio Backyard Billionaires during the SonicBang! festival last week. The band was founded in 2002 and consists of guitarist Sophie Hood, keyboard player James Flores and drummer Tony Marino. Since 2005, the trio is accompanied by Tony's sister Megan who creates live visualisations and doubles as the band's sound-engineer. James told us the following story about the first time that Megan was put in charge of mixing their sound during a concert: We arrived late a bit and had little time for setting up, but we got everything working and Megan managed to create a decent mix for the venue. We got right into our set and played two or three songs, but then we suddenly couldn't hear Sophie anymore.

P From the corner of my eye I saw her1 double-checking her$_2$ gear while we$_3$ continued playing.

Y At the same time I saw Megan panicking at the back of the room.

A She$_1$ was pushing all kinds of buttons trying to get the guitar back, but instead she2 only managed to mute me$_3$ and Tony as well.

T It$_1$ got pretty quiet in the room and soon some people in the audience started laughing and others started to boo.

S It$_1$ was rather awkward.

V Sophie finally figured out that her$_1$ amplifier was the problem.

C It$_1$ had stopped working due to a blown fuse.

I Having found the source of the problem, she$_1$ picked up her guitar again.

L She1 unplugged it$_2$ and hooked it$_3$ up to a second input of my$_4$ speaker.

U It$_1$ didn't work directly, but she$_2$ turned some knobs and managed to get it$_3$ back into the mix.

X (End)
In the end it$_1$ didn't sound as good as before, but at least we$_2$ could finish the gig and prevent a total disaster.

The underlined pronouns were classified as follows:

Ambiguous Personal Pronouns
P1, P2, A1, A2, V1, I1, L1, U2
Ambiguous *it* Pronouns
C1, L2, L3
Under-specified *it* Pronouns
U1, U3, X1
Non-referring *it* Pronouns
T1, S1