

## Disagreement Dissected: Vagueness as a Source of Ambiguity in Nominal (Co-)Reference

Yannick Versley

Seminar für Sprachwissenschaft

Universität Tübingen

E-mail: `versley [at] sfs.uni-tuebingen.de`

Using a qualitative analysis of disagreements from a referentially annotated newspaper corpus, we show that, in coreference annotation, vague referents are prone to greater disagreement. We show how potentially problematic cases can be dealt with in a way that is practical even for larger-scale annotation, considering a real-world example from newspaper text.

### 1 Introduction

Since the early investigations by Hirschman et al. (1997) and the critique of the MUC-7 annotation scheme put forward by van Deemter and Kibble (2000), several large corpora have been annotated with coreference relations, with refinements in terms of annotation schemes (Poesio, 2004), as well as in terms of support by the annotation tools.

After van Deemter and Kibble and their critique of coreference in the case of bound anaphora, critique of the concept of coreference itself came only from Poesio and Reyle (2001), who argue that in the case of mereologically structured entities (both physical objects and abstract objects like plans), it is possible that underspecified references occur without any loss of understandability on the part of the reader/listener, and propose an underspecified DRT representation to cope with these cases.

Poesio and Artstein (2005) argue that coreference annotation can (and possibly should) use underspecified representations to cope with these ambiguous cases. In a study with 18 subjects in an annotation setting, they found that, in the text used, 42.6% of the markables were at least implicitly ambiguous (in the sense that at least two antecedents were chosen by more than two coders each), of which a little more than one third were marked explicitly by annotators when they had been told to do so.

These results are also highly relevant for large-scale annotation efforts like ACE<sup>1</sup>, the dutch KNACK-2002 corpus (Hoste and Daelemans, 2004), or the ongoing effort to add coreference annotation to the text of the German TüBa-D/Z treebank (Hinrichs et al., 2005), since these ambiguities may well occur not only in spoken dialogues but also in edited newspaper texts.

In the remaining part of this paper, we will provide a qualitative analysis of disagreements in the TüBa-D/Z corpus, to show that a certain class of cases, namely those involving vague objects, are prone to genuine ambiguities and lead to a decrease in annotator agreement where they occur. This class includes, but is not limited to, the cases that Poesio and Reyle call the ‘meronymy pattern’.

---

<sup>1</sup><http://projects.ldc.upenn.edu/ace/>

We contend that ambiguities are interesting, even given that we do not wish to annotate these ambiguities explicitly, since we can raise the annotator’s awareness of these ambiguities and propose a resolution of these ambiguities based on independent principles<sup>2</sup>.

## 2 Data examined

In the referentially annotated TüBa-D/Z corpus of written German, we examined the nominal coreference annotations for the 60 articles that had been annotated by two annotators. Inter-annotator agreement (as indicated by F-measure following Vilain et al. (1995)’s scoring scheme) is at 0.83 for all mentions<sup>3</sup>. After normalizing differing spans by mapping them to nodes in the treebank using fuzzy matching, and projecting every markable to the span where it should be following the annotation guide, the inter-annotator agreement improves to F=0.85, which is a visible improvement but less than what Hirschman et al. (1997) found in their study when they let annotators discuss and agree on markables and their boundaries. For full NP mentions only (excluding NPs in predicative positions, as in ‘Peter is [the greatest fool of all]’), the agreement is at F=0.78.

We classified every full NP mention that any one of the two annotators had annotated as being coreferent with another mention (including pronominal mentions) with a semantic class label using the following categorization<sup>4</sup>:

- Persons (PER) are natural persons, including plural person NPs used metonymously to denote some organization (the conservatives, the policemen).
- Organizations (ORG) are formal groupings of persons that are seen as a single actor (e.g. political parties, sports clubs, research institutes)
- Events (EVT) are abstract objects that have a (more or less well-defined) temporal boundary and often result in a change in the state of affairs (e.g. bombings, financial mergers, strikes).
- Locations (LOC) are all geopolitical entities (countries, cities etc.) as well as geographical and physical features.
- Objects (Obj) are things that can be possessed and used and which are generally not seen as being able to perform actions of their own. They may or may not have a material form (as in bank accounts, or electronic books).
- Temporal entities (TMP) are regions of time that are referred to explicitly (e.g. the next week, the eighth day of the strike, Christmas 2006).
- All the rest (Other) is a disconcerting hodge-podge comprising propositions, organizational roles (as opposed to the person filling that role), concepts, legal rights, plans etc. that we did not want to distinguish further.

---

<sup>2</sup>As an example for such independent principles, consider ambiguous modifier attachment in syntax annotation, where ambiguities are usually solved by attaching to the higher candidate.

<sup>3</sup>Hirschman et al. (1997) also give an agreement figure of F=0.83, but they counted the elements of appositional constructions as two markables linked by a coreference relation while we count appositional constructions as a single markable. Because these additional links between appositions are trivial to annotate, the agreement on the remaining relations is probably slightly better in TüBa-D/Z than in MUC-7.

<sup>4</sup>The annotation of semantic classes was performed by the author of this paper. Zaenen et al. (2004), who did a study with 3 annotators for a slightly finer coding scheme, found that the agreement they got for this task was quite good ( $\kappa = 0.92$ ).

	(total)			(disagree)			error rate		
	pl	sg	all	pl	sg	all	pl	sg	all
PER	156	297	453	33	25	58	0.21	0.08	0.13
ORG	38	310	348	8	39	47	0.21	0.13	0.14
LOC	12	204	217	1	41	42	0.08	0.20	0.19
EVT	31	165	196	5	42	47	0.16	0.26	0.24
Obj	29	80	109	11	11	22	0.38	0.14	0.20
TMP	—	14	14	—	5	5	—	0.36	0.36
Other	16	95	111	1	18	19	0.06	0.19	0.17

Table 1: Disagreements by number (singular/plural)

Looking at the disagreement shown in table 1, we can see that there is significant interaction between disagreements and semantic classes ( $\chi^2 = 20.77$ ,  $p < 0.01$ ), and between disagreements and number ( $\chi^2 = 4.76$ ,  $p < 0.05$ ). Single persons, organizations and objects have the lowest error rates<sup>5</sup>, whereas plural objects and temporal entities (which only occurred with singular number) exhibit an unusually high error rate.

Several error types contribute to these discrepancies. If we distinguish between ‘hard’ disagreements, where both annotators’ versions can be seen as equally valid, and soft errors, where it is obvious that one of the annotators just overlooked a possible antecedent, we find that many of the errors for single locations and all of the errors for temporal mentions are indeed soft errors and would possibly profit highly from better annotation tools: in these cases, the location or the temporal region is uniquely (and thus unambiguously) specified, but since they are always uniquely specified (and never anaphoric in the sense that context information from a specific antecedent was needed for the interpretation). Additionally, keeping track of the temporal and spatial locations in a story is usually not required, while keeping track of the protagonists of a story (usually persons and/or organizations) is required for its understanding. In three of the five erroneous coreference decisions regarding temporal mentions, we found that, to realize the coreference relations between the mentions, it would be necessary to make certain inferences that a cursory reader will almost certainly not make.

For plural objects, another source of disagreement is overrepresented, the ambiguity whether a given mention is used in a specific or in a generic sense, typically when a class of objects is denoted. As a simplified example, the sentence “John threw out the red shirts” can have a specific reading (where John acts on a well-defined set of shirts, moving them from his closet to the dustbin) and a generic one (where John expresses his attitude toward the class of red shirts, and he’s less likely to buy one again).

The ‘hard’ disagreements in the PER class mostly involve groups of persons, which are *not* generic, but the actual set of persons that they denote is vague, and annotators decided differently on the question whether two vague objects corefer. Problems with vague reference are usually suspected with event coreference, which is why general event coreference is usually excluded from annotation schemes that are geared towards the reliable annotation of large text quantities, but the presence of this problem for groups of persons (and organizations) suggests that a principled treatment of vague reference would benefit not only the coreference annotation for nominalized events, but also that for groups of persons, which are as important (disagreement-wise) as the former.

<sup>5</sup>We defined the error rate as the ratio between the number of disagreements and the number of markables that were coreferent to another markable in at least one annotator’s version.

### 3 Coreference of Vague Objects

In order to be able to notice, discuss, and possibly resolve ambiguities (or equivalently, argue that a certain annotation is right, wrong, or left ambiguous by the annotation guidelines), we need to complement our naïve understanding with a (semi-)formal description of what coreference is; it is mostly uncontroversial that we build some kind of model from the text, with mentions referring to entities that either have been mentioned in the text or can be accommodated in the model. In terms of the scene that the text describes, it is unlikely that several blatantly dissimilar interpretations arise for the kind of text that we are investigating. Thus, ambiguities must be attributed to the reference relation between mentions on one hand and pieces of modeled reality on the other hand, and identity conditions between these pieces, which are both non-issues with concrete referents, or some vague referents like mountains that can be individuated by their peak. Without an obvious individuation criterion, coreference decisions can become difficult.

Consider the following sentences, taken from the TüBa-D/Z corpus<sup>6</sup>:

- (1) a. Für ein “barrierefreies Bremen” gingen deshalb gestern [1 mehrere hundert behinderte Menschen] auf die Straße – und demonstrierten für “Gleichstellung statt Barrieren”.
- b. “Warum immer wir?” fragten [2 die Versammelten] deshalb auf Plakaten.

It is intuitively clear that the person groups from mentions 1 and 2 have a large overlap, but, seen in isolation, the real-world extensions of the two mentions do not seem to be identical, as not every demonstrator had disabilities, and neither did every one of them carry a poster with the indicated question. On the other hand, saying that 1 and 2 denote different entities would miss the point, since the author meant to talk about the group of demonstrators and not several largely overlapping subsets of it.

We would like to treat the demonstrators as one entity that is described by several predications and not several distinct entities, just as we would not want to talk about multiple clouds when there is just one cloud in the sky to which several predicates apply differently on different parts.

If we treat the conditions of being disabled and of carrying posters as incidental and instead use the demonstrating as the defining property of the crowd of mentions 1 and 2, we can coerce the individual predicates of being disabled, and wanting to push for a “barrier-free Bremen”, to (vague) predicates of groups by taking a majority view.

That is, the article talks about a crowd of demonstrators that

- wanted to push for a “barrier-free Bremen”
- comprised (about) several hundred people
- consisted (in a significant proportion) of disabled people
- had some posters asking “Why always us?” (cumulative reading)

Intuitively, this is much closer to the intended interpretation than talking about several overlapping but not identical crowds. But we have to make sure that we will not run into problems this way, or at least that we know it when we do — if we point to a crowd, it is unclear whether we mean this set of persons or another one that differs in only one person belonging or not to the set, giving us multiple equally possible extensions for that crowd.

---

<sup>6</sup>Translation: (a) For a “barrier-free Bremen”, [1 several hundred disabled people] went onto the streets yesterday — and demonstrated for “Equality, not Barriers”. (b) “Why always us?” [2 the congregated] asked on the posters.

The problem of vagueness in reference has been studied extensively (see Weatherston, 2005), and we will use Smith and Brogaard (2001)'s superevaluationist account of reference to vague objects and predications of these objects.

Smith and Brogaard posit that you can, for a vague object, give multiple precisifications relevant to a certain context – for a cloud, several cloud-shaped sets of water molecules, for a crowd, multiple sets of persons, or, for a lorry loaded with oranges, the lorry with or without the oranges.

A statement is then judgeable and true (supertrue) iff we can instantiate every singular term with a corresponding family of aggregates and that, however we select a single possibility from the family of aggregates, the statement is true.

If we construct a logical form out of the sentences from the crowd example and model all possible crowd extensions, the conclusion (“the crowd from sentence (1) is the same as the crowd from sentence (2)”) will obviously *not* be supertrue since we could always choose two different extensions for the two crowds. Saying that two crowds are the same when they have a large overlap would partially solve the problem, but leads us to Sorites-style paradoxes where the crowd of demonstrators is the same crowd as another crowd etc., where the last crowd of the chain is the same as a crowd totally different from the first one. We can posit identity independently of the extension for objects where we have an individuation criterion, for example the peak for a mountain, or for humans, but not for crowds.

This is where the idea of a (cognitive) model comes in, since we can conceptually separate the process of model construction from the process of model verification (i.e. seeing if the model fits the real world, or asking questions about possible conclusions). This is also done in SDRT (Asher and Lascarides, 2003), where the construction of the discourse model is done using a quantifier-free default logic, whereas the semantics of the model itself uses a monotonic logic without quantifiers. In a similar fashion, we want to handle vague predications in the semantics of the discourse model itself, but not in the construction of the model.

For our crowd example, we could construct a discourse model with a single referent for both mentions and supplement it with possible extensions of the crowd, of consisting of several hundred people etc. such that the predications of the text (in the form of the discourse model, with the given sets of possible extensions) are supertrue. But we could also construct a discourse model with two separate referents for the two mentions. What keeps us from positing another model with two (or even more) overlapping but unrelated crowds?

We could argue that there is nothing that keeps us from positing a model with two overlapping crowds that are both mentioned in the text, and that the distinction between the one-crowd and the two-crowd model is best left underspecified. But we would like to be able to choose as the preferred interpretation one of the two possible discourse models (which we posit can both be constructed from the text and are both supertrue when evaluated in conjunction with plausible predicate extensions).

In terms of the number of entities involved, a model with multiple overlapping crowds would be larger, and by positing more identity relations we decrease the number of entities in the model. We can say that we only want to consider *minimal* models, as proposed by Gardent and Webber (2001), more specifically what they call locally minimal.

This leaves open the question what to do when there are multiple minimal models that all make the statements of the discourse judgeable and true, and it can be argued that an approach using underspecification like the one of Poesio and Reyle (2001) is still needed for these cases. Poesio and Reyle's example of “hooking up the engine to the boxcar and sending *it* to London” is not disambiguated by our criteria since we cannot distinguish between the interpretation where the train is referenced as a (vague) extension of the boxcar and that where

it is referenced as a (vague) extension of the engine, at least not using domain-independent principles. But our account correctly predicts that “stirring up the butter and the sugar and cooking *it* on a stove” is awkward, since the mixed substance cannot be seen as a vague extension of either of the two.

We also hope that using underspecification, or ad-hoc ambiguity resolution, is needed for fewer cases and can be used with greater confidence, allowing for a better compromise between simplicity and annotation quality than relying on multiple annotators to make consistent ad-hoc judgements.

## 4 Conclusion

We have provided a quantitative analysis of disagreements in a referentially annotated corpus, the TüBa-D/Z corpus of written German, and shown that entities with vague extensions like groups of persons are subject to greater-than-average disagreement among annotators. We proposed a theoretical framework based on Smith and Brogaard (2001)’s superevaluationist account of reference to vague objects and minimal models that can help to better understand and resolve difficult cases of coreference, complementing Poesio and Reyle (2001)’s approach by stating further conditions on when underspecification is really necessary. A further study is needed to show if and by how much the improved theoretical framework leads to better agreement among annotators and, generally, better annotation quality, as we hope and at least van Deemter and Kibble (2000) seem to imply in their article.

**Acknowledgements** I would like to thank Heike Zinsmeister, Mareile Knees, Sandra Kübler and Piklu Gupta as well as three anonymous reviewers for helpful comments on a draft of this paper. The author’s work was supported as part of the DFG collaborative research centre (Sonderforschungsbereich) “SFB 441: Linguistische Datenstrukturen”.

## References

- Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.
- Gardent, C. and Webber, B. (2001). Towards the use of automated reasoning in discourse disambiguation. *Journal of Logic, Language and Information*, 10:487–509.
- Hinrichs, E., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor.
- Hirschman, L., Robinson, P., Burger, J., and Vilain, M. (1997). Automating coreference: The role of automated training data. In *Proc. of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.
- Hoste, V. and Daelemans, W. (2004). Learning Dutch coreference resolution. In *Fifteenth Computational Linguistics in the Netherlands Meeting (CLIN 2004)*.
- Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL’04*, Boston.
- Poesio, M. and Artstein, R. (2005). Annotating (anaphoric) ambiguity. In *Corpus Linguistics 2005*, Birmingham.
- Poesio, M. and Reyle, U. (2001). Underspecification in anaphoric reference. In *Fourth International Workshop on Computational Semantics (IWCS-4)*.
- Smith, B. and Brogaard, B. (2001). A unified theory of truth and reference. *Logique et Analyse*, 43(169-170):49–93.

- van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.
- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann.
- Weatherson, B. (2005). The problem of the many. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University.
- Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitana, T., O'Connor, M. C., and Wasow, T. (2004). Animacy encoding in english: why and how. In *ACL 2004 Workshop on Discourse Annotation*.