Vagueness and Referential Ambiguity in a Large-scale Annotated Corpus*

Yannick Versley

January 2008

Abstract

In this paper, we argue that difficulties in the definition of coreference itself contribute to lower inter-annotator agreement in certain cases. Data from a large referentially annotated corpus serves to corroborate this point, using a quantitative investigation to assess which effects or problems are likely to be the most prominent. Several examples where such problems occur are discussed in more detail, and we then propose a generalisation of Poesio, Reyle and Stevenson's Justified Sloppiness Hypothesis to provide a unified model for these cases of disagreement and argue that a deeper understanding of the phenomena involved allows to tackle problematic cases in a more principled fashion than would be possible using only pre-theoretic intuitions.

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.

Creating large annotated corpora like the ACE corpus[1], the Dutch KNACK-2002 corpus (Hoste and Daelemans, 2004), or the referential layer of the German TüBa-D/Z treebank (Hinrichs et al., 2005) entails a significant cost in terms of the man-power needed both for the annotation and for the development of annotation guidelines and, possibly, annotation tools. This cost is usually justified by data-driven development of systems that automatically do coreference resolution: a vast majority of the systems that are now considered state of the art (Luo et al., 2004, Uryupina, 2006, *inter alia*) use models that are automatically learned from a corpus.

[*] In Massimo Poesio and Ron Artstein (eds.): Ambiguity in Anaphora. Special Issue of the Journal on Research on Language and Computation, *to appear*.

[1] `http://projects.ldc.upenn.edu/ace/`

Looking at a neighbouring field, we see that annotated corpora can remain useful for a very long time, even when the framework used for building the systems varies considerably: the Penn Treebank has been in use for the data-driven building of parsers for quite a long time, from (Magerman, 1995) to recent approaches like (Charniak and Johnson, 2005), who use different features and get qualitatively better results. There is even research that uses the Penn Treebank data to induce probabilistic grammars that yield linguistically richer structures than directly available in the Penn Treebank (Hockenmaier and Steedman, 2002; Cahill et al., 2002; Miyao and Tsujii, 2005).

But data-driven system development is not the only use of annotated corpora, and nor should it be. Annotated corpora can be very useful for corpus linguistic research, as well as a source of examples for research that normally tends to rely on introspection. Especially the investigation of rare phenomena can profit from annotated corpora since it is easier to pinpoint possible occurrences than with approximate queries in unannotated corpora (Meurers, 2005).

The quality of corpus annotation is important both for data-driven systems development and for corpus-linguistic research: for data-driven systems development, repeatability (or consistency) of annotation obviously yields an upper limit to what an automatic system can do; and even when machine learning approaches can cope with noisy data quite well, they do better when less additional noise is present. Dickinson and Meurers (2005) show that parsing performance improves visibly when using a corrected version of the Penn Treebank for training. Corpus linguists do also profit from high annotation quality, since incorrectly annotated data leads to more spurious matches as well as potentially interesting items that are overlooked because they are annotated incorrectly.

In the following section, we will give an overview of common methods for assessing annotation quality, explain the notion of referential ambiguities as annotation problems arising from a mismatch between annotation model and the text meaning it is meant to model. We try to shed some light on potentially problematic issues in coreference resolution by outlining some criteria where we think coreference annotation is *less* problematic, namely the resolution of (pronominal) anaphora to textual antecedents, which is linguistically well understood, and the tracking of entities coming from a fixed frame of reference (named entity tracking), where reference relations are ontologically well-founded; whereas coreference annotation also encompasses other phenomena that are less well-founded and that should be subject to more research.

In section 3, we will make the case that referential ambiguities exist not only in spontaneous dialogue such as in the corpus examined by Poesio and Artstein (2005), but also in large referentially annotated corpora of written language, such as TüBa-D/Z. Using a quantitative analysis, we will also argue that there is a marked difference in the achieved consistency between annotation that fulfills the criteria outlined earlier, which can be annotated very consistently, and annotation where these criteria are not met. We will then analyse some examples where problems arise from these assumptions not being met in more detail, in section 4, and, finally, in section 5, come back to the question of how to treat referential ambiguity, comparing the merits of several proposed solutions.

2 Annotation Quality

There are two criteria that are important when assessing a given annotation: One is that the annotation makes explicit some meaning that is inherent in the data but would be hard to get at otherwise. Although this criterion is very important (it is the reason to do annotation in the first place), meaning is something that is often hard to pin down, and in this respect the (often pre-theoretic) formalisation that is done for the purpose of creating annotation guidelines is no different from other formalisation attempts in the framework of particular theories.

The other criterion, which is more easily verified, is repeatability, or consistency of annotations. This means that someone who knows the guidelines will be able to predict the annotation for a certain phenomenon he or she has in mind; this is important both for corpus linguistic purposes, where we want to search for a particular phenomenon, and for the use in data-driven systems development, where the prediction is to be made automatically and inconsistencies in the annotation limit its usefulness for evaluation and/or model induction purposes.

For coreference annotation or any other annotation task that includes complex structure, one does well to take this complexity into account when designing the workflow leading to the final annotated corpus. In a pilot study on the annotation of temporal relations in texts, Setzer and Gaizauskas (2001) found that several sources of errors for the process of identifying temporal relations between events were related to the complexity of the task: lack of understanding of the guidelines on the part of the annotators (maybe due to insufficient training), fatigue due to the monotony of (parts of) the annotation task or simply carelessness on the part of annotators. These factors make it obvious that with increasing complexity of the annotation task, both extensive training of annotators and good tool support is necessary, but also that a subsequent revision step should be considered a regular part of the annotation process.

Simple mistakes on the part of annotators can be later found and corrected, either by a process of careful cross-checking of annotations (often by a more experienced person), or from error reports by users of the annotated corpus; these can then be verified against the annotation guidelines and eventually the annotation will converge towards a state where it is virtually free of mistakes.

This is different from cases that are not decidable on the basis of the annotation guidelines and the text at hand, where there is (as Setzer and Gaizauskas call it) an "intrinsic difficulty" in determining the correct annotation, due to genuine ambiguities. In the presence of such ambiguities, it is usually difficult to argue which of the alternatives would be correct; moreover, Poesio and Artstein (2005) found that some ambiguities were not even perceived as such when annotators were told to explicitly mark them, which means that a subsequent process of revision may or may not converge to a stable point, and that users of the annotated corpus may not find what they expect and, even in full knowledge of annotation guidelines, consider these annotations to be erroneous.

Some ambiguities, which Knees (2006) calls *structural ambiguities*, only pertain to non-essential information that does not contribute to the meaning conveyed by the annotation, for example the exact text spans of markable in referential or discourse annotation. It is not only possible, but also desirable to eliminate these ambiguities by arbitrarily preferring one solution over the other and fixing this in the annotation guidelines (for example, always taking the largest possible or the smallest possible span for a markable).

In cases where categories posited by the annotation scheme overlap, other structural ambiguities arise, for example between unique descriptions that can be interpreted out of context (called *autonome* in the coreference scheme of Gardent and Manuélian, 2005) and discourse-old descriptions (called *coréferentielles* in their scheme). In cases where a discourse-old unique description is mentioned, Gardent and Manuélian specify that the unique description label is to take precedence. Chiarcos and Krasavina (2005) also make extensive use of such precedence rules to eliminate most cases of structural ambiguity.

The other cases, subsumed by Knees under the label *referential ambiguities*, do pertain to information that is essential to the annotation purpose, and one has to be very careful in crafting annotation guidelines that will both ensure reasonably consistent annotation and also minimise the distortion of the relation between the annotation and its intended meaning, including generalisations and/or predictions that may be only implicit in the annotation scheme.

2.1 Quantitative measures of Annotation quality

The most commonly used method to assess the quality of annotation is to let several annotators independently annotate the same text(s), thus providing evidence for the repeatability of the annotation task (or lack thereof). Since the real annotation process includes an additional step with a more experienced expert correcting many of simple mistakes (whereas ambiguities would still persist), such agreement studies provide some indications of the final quality, but should by no means be considered as predictive for the final quality of the annotation.[2]

For the quantitative analysis of agreement data, it is usual to use some statistic that is monotonic with agreement (i.e., gives a number that increases when more cases are agreed upon) and lies in a range that is independent of document/corpus size and possibly other incidental properties of the annotated subcorpus.

The simplest statistic, percent agreement (i.e. number of cases agreed upon

[2] An example for this comes the SALSA corpus (Burchardt et al., 2006), where FrameNet annotations are added to a treebank: in addition to inter-annotator agreement, they calculate agreement between adjudicators, who create consensus for about 80% of the disagreements. Instead of the 20% of cases where the original annotators did not agree, adjudication thus leaves about 4% that are truly difficult cases.

divided by total number of cases) has two obvious problems: First, it is not readily applicable to annotation that goes beyond assigning a category to single items. Second, it is sensitive to frequency bias in that a single large class can make raw agreement artificially high (consider two psychologists judging 10% of a sample as mentally ill and agreeing on 90% of the - mainly non-ill - cases purely by chance).

The latter problem, uneven distribution of classes, has been approached with two different solutions. The first, traditionally used in the social sciences and popularised for quantitative agreement studies in linguistic annotation by Carletta (1996), are chance-corrected agreement measures such as the kappa coefficient introduced by Cohen (1960).

Assuming two annotators X and Y, we call p_{xy} the probability that something is marked by both X and Y, $p_{x\overline{y}}$ the probability that something is marked by X, but not by Y, etc., the observed percent agreement would be $A_o := p_{xy} + p_{\overline{xy}}$, whereas the agreement expected for random annotation would be $A_e := p_x p_y + p_{\overline{x}} p_{\overline{y}}$ (where p_x is the probability that something is marked by X). The kappa coefficient maps the range from expected (A_e) to perfect (100%) agreement to the interval $[0, 1]$:

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

The other solution, F-measure, also called *positive specific agreement*, comes from Information Retrieval (van Rijsbergen, 1979) and works well when one class (relevant documents) is much smaller than the other class (irrelevant ones). Focusing only on the minority class of positive instances, it is possible to say how many positive instances are agreed upon as a fraction of the total positive instances for one annotator, which we will call $P = \frac{p_{xy}}{p_x}$ and $R = \frac{p_{xy}}{p_y}$.[3] Taking the harmonic mean of these proportions, we get

$$F = \frac{2PR}{P + R} = \frac{2p_{xy}}{p_x + p_y}$$

Vilain et al. (1995) have proposed a model-theoretic adaptation of F that extends the related concepts to partitionings. Based on Vilain et al.'s model, Passonneau (1997) proposes a kappa agreement measure by considering absent links between entities (or entity groups) as negative class instances and computing an expectation for the joint agreement probability on positive and negative cases.

2.2 Defining Coreference

In the time since McCarthy and Lehnert (1995) argued for the utility of coreference classification in information extraction (which was, and partly still is, dominated by

[3] Assuming random annotation, we would get $P = \frac{p_{xy}}{p_x} = p_y$ by independence assumption, and random agreement would be close to zero if p_y (and for the converse case, p_x) are, i.e., if the positive class is small. While originally conceived as an evaluation measure and not as an agreement measure, some researchers recommend to use F alongside κ (Cicchetti and Feinstein, 1990), or in cases where the size of the negative class is very large (Hripcsak and Rothschild, 2005).

more shallow processing and for the most part avoids relying on richer linguistical representations), and formal annotation guidelines were drafted for the MUC-6 coreference task (MUC6, 1995), the understanding of coreference annotation has been greatly improved by criticisms based on theoretical consideration, and, on the other hand, the theoretical understanding of the underlying mechanisms and representations of coreference has benefitted from the additional attention brought by annotation and annotation problems.

The initial definition of coreference proposed for the MUC-6 coreference task was very broad, also encompassing a few cases that would now be considered identity-of-sense anaphora (the aluminium market → aluminium) and not coreferent in the proper sense. Furthermore, van Deemter and Kibble (2000) pointed out that a number of problems exist with the MUC definitions: using the currently assumed notion of reference with respect to the real world, quantified expressions and pronouns that refer to them do not have a (unique, well-defined) referent in the real world, and should not be included in definitions of coreference. They also point out problems with predicative constructions, where the predicate does is not a referring expression, but is something that is predicted. (For example, in the sentence "The stock price rose from $4.56 to $5.02", both prices are something that is predicted, but neither is coreferent with the NP "the stock price" in the stricter sense).

Van Deemter and Kibble also point out that coreference, seen as a relation from textual entities to entities in the world as in the MUC guidelines, becomes problematic for quantified expression (like "every student who eats a big Kahuna burger enjoys it"), since the quantified NP does not refer to a single referent (and nor does the pronoun, which is clearly anaphoric).

Karttunen (1976), on the other hand, starts with the idea of coindexation that is used in binding theory and proposes to extend this to pronominal anaphora generally and demonstrates some rules that govern the interaction between pronominal reference on one hand and quantification or modal embedding on the other. For example, in "Peter wants to buy a new car. *It* would have to be red.", the anaphoric description '*it*' is perfectly acceptable and there is no problem resolving it to its antecedent, even when there is no actual real-world car involved. Most existing theories of discourse such as (S)DRT (Asher and Lascarides, 2003) assume anaphoric reference to happen as coindexation of entities in a model (discourse referents), where the discourse referents in a dynamic logic setting are semantically not different from bound variables in logical formulae when it comes to the representation of anaphora bound by quantifiers.

Indeed, restricting ourselves to concrete real-world entities would imply that anaphora annotation in biomedical texts be all but meaningless, since there, the relation between textual entities and real-world entities is not at all straightforward. But consider the following example (taken from Castaño et al., 2002):

(1) We quantitatively analyzed the relationship between the structure and inhibiting activity of these substances toward [acetylcholinesterase and bu-

> tyryl cholinesterase]. Hydrophobic interactions were found to be important for the inhibition of [both enzymes] but are more pronounced in the case of butyryl chlinesterase.

The NP “acetylcholinesterase” does not designate a concrete molecule (or chunk of molecules) in the world but the text makes a general prediction about the mentioned enzymes. Nonetheless, the anaphoric reference by the NP “both enzymes” is not at all problematic, since we have no problem of finding a textual antecedent.

Thus, we need to replace the concept of reference to entities in the real word by something more adequate; a good first approximation would be to speak of reference to entities in a mental model. Mental models can (potentially) be somewhat idiosyncratic; biomedical texts nonetheless have precise reference properties because there is a frame of reference underlying them that is shared by all users of the (sub)language at hand. More generally, we can predict that reference (and coreference) is unproblematic as long as it is textually unambiguous (as in the case of pronominal anaphora, bound or not) or there is a common frame of reference containing a discrete set of non-overlapping entities. The existence of such a frame of reference is usually a precondition to the use of rigid designators (proper names, but also other names such as the substance names in biochemical texts or medical leaflets, which the GNOME annotation treats very similarly to proper names, cf. Poesio, 2000), and it ensures that properties that are ascribed to the entity in one piece of the text are compatible with descriptions in any other piece of the same text (substitutability).

3 Quantitative Evidence for Problematic Cases

For the following analysis, we used cases of disagreements found in two portions of the current release of the Tübingen Treebank of Written German (TüBa-D/Z, Hinrichs et al., 2005) that have been annotated by multiple annotators. A first portion, totalling about 60 articles, was annotated by two annotators, and the second, totalling about 80 articles, was annotated by three annotators. One annotator, who annotated both portions, was a post-graduate student of computational linguistics. The other three annotators were undergraduate students of computational or general linguistics.

Inter-annotator agreement on coreference sets (as F-measure following Vilain et al.’s scoring scheme) is at 0.83 for the first portion, and varies between 0.82 and 0.84 for the second portion.[4] If we automatically enforce the requirement to include *all* modifiers in the markable span using the existing syntactic annotation, we get an improved inter-annotator agreement of F=0.85 (for the second portion, between 0.85 and 0.89). This is a visible improvement, but less than what Hirschman

[4] 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 them 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.

type	F	κ
anaphoric	0.95	0.86
coref-NE	0.89	0.84
coref-nonNE	0.69	0.65
reflexive	0.72	0.58
relative	0.92	0.30
ALL	0.85	0.81

Table 1: Agreement by link type

et al. (1997) found in their study when they let annotators discuss and agree on markables and their boundaries.

We used an adaptation of Passoneau's (1997) scoring scheme[5] that allows us to get κ and F-based agreement measures for subsets of the coreference links to assess the reliability not only for the annotation task as a whole, but for several parts of the task that we think are different enough to be evaluated separately (see table 1). We split the markables according to (personal and demonstrative) pronouns (*anaphoric*), full noun phrase coreference with a named antecedent (*coref-NE*), full noun phrase coreference with a non-named antecedent (*coref-nonNE*), as well as reflexive and relative pronouns.

The overall κ value is 0.81, which counts as good reliability in the ranges cited by Di Eugenio and Glass (2004). Parallel to the predictions made earlier, both those coreference links between pronominal anaphora and their antecedents ($F = 0.96$, $\kappa = 0.86$) and coreference links to named entities ($F = 0.89$, $\kappa = 0.84$) can be annotated very reliably, but coreference between arbitrary common nouns is somewhat problematic ($F = 0.69$, $\kappa = 0.65$). The annotation of reflexive pronouns was subject to a requirement that non-referring reflexive pronouns (occurring with so-called inherently reflexive verbs, which always require a reflexive pronoun as object) are not to be annotated. The distinction between normal verbs with a reflexive pronoun and inherently reflexive verbs seems to be somewhat problematic, something which is reflected in lower F and κ scores. Relative pronouns can be annotated quite reliably (as witnessed by the high F value), but since they always have a referent and the distribution is very skewed towards annotating them, the κ value does not reflect this. We see this latter effect as a flaw in the kappa statistic used rather than as a serious annotation problem.

In an earlier study, we classified every full NP mention in the first portion 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[6]:

[5] We derive a maximally consensual directed spanning tree of links and classify each markable as agreeing(no-link), agreeing(same-antecedent) or disagreeing, then proceeding as proposed by Passonneau.

[6] 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

- 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) have a (more or less well-defined) temporal boundary and often result in a change in the state of affairs (e.g. wars, 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. 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).
- The rest (OTHER) is a mixture comprising abstract objects that we did not want to distinguish further, including propositions, legal rights, plans etc.

Looking at the disagreement shown in table 2, 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[7], whereas plural objects and temporal entities (which only occured with singular number) exhibit an unusually high error rate.

Several error types contribute to these discrepancies. We find that many of the errors for single locations and all of the errors for temporal mentions are indeed simply mistakes 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, rather than being 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

they got for this task was quite good ($\kappa = 0.92$). Unfortunately, non-repeated noun phrase mentions have not been annotated with semantic class information, which precludes the application of κ-based measures of per-class agreement on the coreference annotation.

[7] 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.

	(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 2: Disagreements by semantic class and number

sentence “They want to make the freight trains faster” can have a specific reading (where the trains get faster individually) and a generic one (where slow trains are decommissioned and replaced by faster ones, such that the [average] speed of the freight train fleet increases, even though no train is actually made faster).

The disagreement in the PER class that were due to ambiguities 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.

4 Instances of Referential Ambiguity

In the last section, we provided quantitative evidence to support our hypothesis that coreference can be annotated very reliably as long as it is a textual phenomenon (anaphora) or the mentions in the text refer to non-overlapping entities from a fixed frame of reference (named entity tracking), but can be problematic in other cases.

More specifically, we will examine the phenomena that arise when either the entities are not discrete, but have a non-negligible problem of vagueness; when the entities are well-defined but there are multiple frames of reference; or, finally, when mentions evoke (and possibly refer to) more than one entity, as can be the case in polysemy and metonymy but also in other cases.

We use specific examples to examine some of these cases where the posited criteria do not hold and investigate the nature of the resulting problems, possibly making reference to specific theories of discourse for explanation.

4.1 Co-reference of Vague Entities

At least in newspaper text, it is a plausible assumption that people reading a text will get a very similar view of the main events or propositions of said text, even when annotating that text might uncover referential ambiguities. Our working hypothesis would therefore be that referential ambiguities must be attributed to the reference relation between mentions and pieces of modeled reality, as well as the identity conditions between these pieces, which are both non-issues with concrete referents, or some referents like mountains that have vague boundaries but 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[8]:

(2) 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.

If we use both descriptions in isolation to delineate an extension for the group entities that each mention refers to (taking a very literal approach of co-reference as first determining a real-world entity for each mention and then saying that they co-refer if the two entities are the same), it is intuitively clear that the person groups from mentions 1 and 2 must 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, 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)

[8] Translated from TüBa-D/Z; German original:
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.

This model is preferable on the grounds of being *minimal*: for two distinct entities in the model, there is some information (either in the text or from background knowledge) that distinguishes them, and it is preferable to talking about several overlapping but not identical clouds.

But this also means that we have to *first* build a model (including the merging the 'referential indices' of coreferent mentions) and *then* interpret it, as opposed to first taking a relation of mentions to (real-world) entities and then judging coreference on the equality of referents, which will lead to strange results. Conversely, the model-first approach that we want to advocate here weakens the idea of *substitutability* of two textual descriptions to one of compatibility of a description with some discourse entity. Unless we are ready to admit vagueness — multiple possible extensions for a group, even though the definite article suggest that the expression has a unique denotation — we would have to define this compatibility in similar terms than for substitutability, which would still be problematic.

The proposed model would have the denotations for the real-world entities to shift in the process of incremental model-building that accompanies text comprehension, something which is not intuitive but, as we will argue, must be considered a realistic assumption.

To apply the minimal model criterion, we still need to be able to discuss whether merging two entities in the model would lead to a contradiction (assuming commonsense background assumptions), and it might be argued that allowing discourse entities to be described by vague predicates lets the judgements on them, including compatibility or mergeability, become vague, and possibly ambiguous.

To see that this is not the case, let us consider how the problem of vagueness in reference is solved in Smith and Brogaard (2001)'s supervaluationist 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 house that has an annex built besides it, the house with or without the annex.

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.

In our example, the predicates of having a certain political objective (by extension from the members to the crowd), to comprise several hundred people, to consist (in a significant proportion) of disabled people and of carrying some posters (in a cumulative sense) are all vague, and not every precisification of one predicate is a plausible precisification of another predicate. But if we use a conjunction of all predicates (a straightforward way of interpreting this would be to take the intersection of the individual precisifications), we still get a plausible set of precisifications for the whole description. Thus, we can instantiate the term that represents our referent in the model with a family of precisifications that makes the whole description judgeable and true (which is what we wanted in the first place). The set of possible precisifications for the joint descriptions would then be a subset of the precisifications for an individual mention, but the set of *preferred* precisifications could be a different one, thus explaining a shift in the (vague) denotation in the course of incremental model-building (where we first have a description only based on the first mention, which is subsequently enrichted to include the information from the second mention).

4.2 Reference in Blended Spaces

The account of vague extensions presented in the preceding section allows a clearer view on disagreements between a markable being discourse-new and it being discourse-old (and coreferent to an earlier mention), by looking at plausible precisifications for referents where an individuation criterion is not available.

Some other ambiguities are due to a richer structure of the mental model than what can be represented as an equivalence relation. Annotating ambiguities would not do justice to the model structure either, since some ambiguities can be related, as in the following example[9]:

(3) a. As a lawyer in Boston, [1 John Travolta] sues two businesses that he holds responsible for eight children having died of leukemia.
b. At first, [2 the calculating career lawyer] only scents the high amount of compensation (...).
c. A court drama, environmental thriller and great actors' cinema, in which [3 Travolta] and his antagonist Robert Duvall reach top form.

Markables 2 (the career lawyer), the lawyer figure from the story and 3 (Travolta), the actor, are not compatible, since they mention two different entities. But markable 1 mentions John Travolta in a context that obviously belongs to the career lawyer. Annotating markable 2 and 3 as ambiguous between discourse-new and coreferent to markable 1 would miss the point, since the interpretation where 2 and 3 are both discourse-new is not felicitous. In terms of the theory of mental spaces (Fauconnier, 1984), we have two spaces, one with actors and one with the story's protagonists, and a third one where the two are blended together. Based on this account, we can choose to consider the blended space for the equivalence relation that we base coreference on.

The solution of using a blended space can also be cast in terms of the *dot objects* described by Asher and Pustejovsky (2005): To explain the semantics of sentences like "The book was a huge pain to lug home and turned out to be very uninteresting", where both the physical object and the book's content are mentioned, or "Mary read the subway wall", where an inference has to be made that

[9] example translated from TüBa-D/Z corpus, sent. 3772ff;
German original:[1 John Travolta] verklagt als Bostoner Anwalt zwei Firmen, die er für den Leukämietod von acht Kindern verantwortlich macht.
Anfangs wittert [2 der berechnende Karriereanwalt] nur die hohe Entschädigungssumme (...).
Gerichtsdrama, Umweltthriller und großes Schauspielkino, in dem [3 Travolta] und sein Gegenspieler Robert Duvall zu Hochform auflaufen.

yields an *informational* aspect of the subway wall, they introduce compositions of primitive types; the book would then be modeled to be of the type *physical* • *informational*, and would have as aspects both the physical object as well as the information contained in it. Asher (2006) uses this to explain the felicity of sentences like "*Superman always gets more dates than Clark Kent does.*" where the usual denotation of *Superman* and *Clark Kent* would be a single individual. By introducing aspectual objects for *Superman as Superman* and *Superman as Clark Kent* (which behave differently with respect do having dates, depending on whether Superman shows up in his Superman role or as Clark Kent) that are both aspects of the (usual) person Superman.

For our purposes, we could represent the entities in the blended space as *actor* • *role* dot objects which have both an actor aspect (reach top form in a court drama) and a fictional person aspect (sue two businesses). The coreference annotation would then consider this dot object (corresponding to an entity in the blended space) as referent for the coreference chain. This solution fails when we have multiple blended spaces for different film adaptations of the same story, where we would get multiple dot objects with the same role, but different actors, and there would not be one preferred blended space. Such texts are rare enough not to be a problem in practice, but they exist.[10]

4.3 Incompatible precisifications

In some cases, we find that, on a coarse level of detail, some entity is repeatedly mentioned and then taken up with an anaphoric mention, whereas on a finer level of detail, the repeated mentions are incompatible, creating a problem of referential ambiguity[11]:

(4) a. The concepts range from a unilateral cessation of the air strikes (...) to a politico-economic Marshall plan for [1 the whole Balkan].

b. At the beginning of the week, US ambassador Robert Barry has (...) summarized his thoughts regarding a sustainable civil solution for [2 the permanent problem region of ex-Yugoslavia].

c. Barry demands (...) a restructuring of the development aid measures in [3 the whole region].

[10] Consider the comparison of different adaptations of the Dickens novel "Great Expectations", found at http://www.victorianweb.org/authors/dickens/ge/filmadapt.html: whereas in one adaptation, "Pip rips down the draperies of Satis House to let in the light of day upon the mould and decay of Satis House and release Estella", another is showing "Pip's snobbery", all while the reader is perfectly comfortable with this over-abundance of blended spaces which would make coreference annotation overly problematic.

[11] translated from TüBa-D/Z sentences 6387,6389,6393;
German original: Die Konzepte reichen von einseitiger Einstellung der Luftangriffe (...) bis zu einem politisch-wirtschaftlichen Marshallplan für [1 den gesamten Balkan].
Zu Beginn der Woche hat US-Botschafter Robert Barry ... seine Gedanken zu einer nachhaltigen zivilen Lösung für [2 die Dauer-Problemregion Ex-Jugoslawien] zusammengefasst.
Barry fordert ... eine Umstrukturierung der Ausbauhilfen in [3 der gesamten Region].

In this case, the theme is a plan by Robert Barry to stabilize the war-riddled region around ex-Yugoslavia by means of targeted financial aid. The exact region isn't specified – and probably the plan is not this detailed yet, so the region is vague and its exact limits are underspecified. But the article mentions this region first as "*the whole Balkan*", and then as "*the (...) region of ex-Yugoslavia*", which are both plausible extensions of this region but not compatible with each other. As a result, one annotator marked markable 1 as coreferent with "the whole region", while the other chose markable 2.

A precondition for this anaphoric reference to a vague entity to be felicitous would be that the reader does not notice the incompatibility between the two previous mentions. From a production perspective, we could argue that the vague entity that the author had in mind was coerced to a more specific one (by the unavailability of more basic terms for the vague region); from a comprehension perspective, the construction-integration model of Kintsch and van Dijk (1978) would predict that some facts from the discourse get simplified in the process of understanding (the sentences have some text between them), and it would be plausible that the more specific representations of "*the whole Balkan*" and "*the (...) region of ex-Yugoslavia*" are reduced to the underlying vague representation that we posit here.

We can formalise this by saying that the mentions evoke both an entity from a space with precise representations, and one from a space with coarser representations, yielding a composite object with a vague and a precise aspect (where the vague aspect is needed for the structuring/comprehension of discourse) – this would not help us immediately for possible solutions since the precise aspect could be referred to in a different context (consider a similar text where "*the (...) region of ex-Yugoslavia*" is followed by mentions of ex-Yugoslavia that elaborate on other propositions and thus don't have the vague region aspect that the first mention has). But we gain at least some explanatory adequacy by the use of this device since we can then argue about the presence of the additional aspect pertaining to a coarser frame of reference.

Positing dot objects with aspects for different levels of granularity (which would correspond to the 'local' theories of different granularity levels, as posited by Hobbs, 1985) would provide an insight into other cases of coreference with two incompatible (potential) antecedents, as in the following example from (Knees, 2006)[12]:

(5) a. It was Robert Jackson's will that this should not remain an isolated case; [At that time already, he demanded that an International Criminal Court should be established].

b. He wanted to take advantage of the favourable moment as the world was shocked by the atrocities of the Nazis, the full enormity of which could only be anticipated first at [Nuremberg]. (...)

c. But an International Criminal Court only starts gradually to take shape

[12] translated from the German original version containing the pronominal adverb *danach*

today, 50 years after [that], in The Hague, where the war crimes of the former Yugoslavia are being dealt with.

Both the Nuremberg trials and Jackson's demanding are plausible antecedents for the anaphoric "after that", but they are obviously *not* compatible. We can model this by positing a coarser granularity level (since the anchoring description *50 years after that* is coarser than the granularity of the anchoring events *the Nuremberg trials* and *Jackson's demanding*) and saying that the anchoring description coerces the antecedent descriptions to a coarser granularity.

Knees remarks that this could be an example for the Justified Sloppiness Hypothesis brought forth in (Poesio et al., 2003), which was motivated by examples like the following, which Poesio and Reyle (2001) found in their study of the TRAINS corpus:

(6) a. Hook up [engine E2] to [the boxcar] at Elmira.
b. And send [it] to Corning as soon as possible.
c. As soon as it arrives, [it] should be filled with oranges.

In this case, annotators disagreed whether the pronoun "it" was referring to the boxcar or the engine. Poesio and Reyle argue that the pronoun (in b) could be seen as referring either to the boxcar, the engine or the whole train (as the mereological sum of both), and argue that a theory of discourse might leave the pronoun interpretation under-specified between all three solutions. In a case like the above including (c), the under-specification gets fully specified later (in c), since only boxcars can be filled with oranges.

Poesio et al. account for this phenomenon with the hypothesis that ambiguity between multiple potential antecedents can occur without being rejected as being ungrammatical (which would certainly occur in the case of an ambiguity that allows for different scenario interpretations, a case where we would see repairs in dialogue or largely negative judgements in a magnitude estimation study), as long as the following conditions are met:

1. Both explicitly mentioned potential antecedents x and y are elements of an underlying mereological structure with summum $\sigma = x \oplus y$ which has been explicitly constructed (and made salient) in the dialogue.

2. the existence of this structure makes it possible to construct a p-underspecified interpretation in which the anaphoric expression is interpreted as denoting an element z included in the mereological structure – i.e. part-of its summum σ:
$$[x\, y\, \sigma\, z \,|\, \ldots \sigma = x \oplus y,\ z \lhd^{*} \sigma \ldots]$$

3. All possible interpretations (x, y, z, $x \oplus y$) are equivalent for the purposes of the plan.

Assuming a mereological hierarchy like that of Reitsma and Bittner (2003)[13], the Justified Sloppiness Hypothesis provides a good explanation for examples like (5) and (6). However, we could not use the Justified Sloppiness Hypothesis to explain our earlier example (4) since the precisifications are not mereological parts of the vague region; also, the reference to mereological structure could be seen as somewhat arbitrary.

In our model, the trains example would be explained by the fact that "*send ? to Corning*" selects for a train and thus triggers the dot-introduction, leading to a *boxcar* • *train* object where only the train aspect is used. Since the other aspect is not used, we can then posit underspecification between a *boxcar* • *train* and a *engine* • *train* entity as referent for the "*it*" in (b). The mereological structure then serves to fulfill the precondition for the introduction of dot objects, but is not special otherwise.

We can formalise this extension of Poesio et al.'s Justified Sloppiness Hypothesis as the following *Generalised Sloppiness Hypothesis*:

Multiple potential antecedents can occur without being rejected as ungrammatical as long as

1. The anaphoric expression occurs as argument to a predicate $P : \beta \multimap t$, i.e. the context selects for a type β.
2. The potential antecedents $x : \alpha_1$, $y : \alpha_2$ have to be extended to a β-compatible complex type (i.e., $\alpha_1 \bullet \beta$ or $\alpha_2 \bullet \beta$).
3. In both cases, the same entity $z : \beta$ is selected as result of the dot introduction.

The notion of dot objects allows us to treat blended spaces, granularity shifts, and cases of polysemy in a uniform way, much like Mani's (1998) treatment of granularity shifts and polysemy using abstraction operators, and it possibly allows us to treat a greater bandwith of phenomena than with Poesio et al's original formulation.

It is also possible to predict the acceptability (or, in Poesio's terms, justification) of such examples as resulting of (i) the plausibility of a coercion from $\alpha_{1/2}$ to β as a result of the context (P) and (ii) the interference from a reading without coercion which would make the example unacceptable.

The difference between coercibility criteria and Poesio et al.'s formulation using plans would be visible in a hypothetical case where the plan is to blow up both the engine and the boxcar by joining them together and planting a bomb on them:

[13] Link (1983)'s original version contained two hierarchies, one of atomic objects and plurals composed of several atomic objects, and one of unstructured lumps, connected by an extension function that maps (groups of) atoms to the lump that is their material extension. Since σ is not a plural and not a simple sum of extensions (as in "the material making up John's left hand and the Empire State Building"), we need to posit a (possibly domain-specific) mereology of functional or otherwise salient parts and wholes.

(7) a. Hook up [engine E2] to [the boxcar] at Elmira.
b. Attach a large bomb to [it] and blow it up.

For the purposes of the plan, it is not important whether the bomb is planted on the engine, the boxcar or the train (assuming that the bomb is large enough), but the predicate of ‘attaching a bomb to something’ does not require its argument to be a train, and no dot introduction is involved. Our Generalized Sloppiness criterion would thus (correctly) predict that (7)b is awkward.

5 Possible treatments for referential ambiguity

As seen in the preceding sections, referential ambiguity is a problem that also occurs in the written texts that are used for large-scale coreference annotation. We think that some (pre-theoretic) notions used to model or to explain coreference contribute to the resulting problems.

Coreference is usually modeled as an *equivalence relation* between entities on the *text surface*, defined by the entities they refer to. In fact, this makes it easy to subsume anaphora under coreference, because of *substitutability*: being able to replace the anaphor with its antecedent would then ensure both its antecedenthood and the coreference to the anaphor. Substitutability, and the other criteria, are obviously met when we consider the case of a discrete set of entities coming from a single, fixed frame of reference (e.g., in the case of most named entities), or most cases of pronominal anaphora.

Already van Deemter and Kibble (2000) raise objections against the conflation of anaphora and coreference, using (non-)substitutability of quantified expressions with further pronominal mentions. If we use coindexation instead of reference to real-world entities and are aware of quantification, van Deemter and Kibble’s point seems much less critical; but the examples in sections 4.1 and 4.2 show that naïve substitutability is indeed not a good criterion for judging coreference. Accordingly, the notion of model building (or mental models more generally) should be seen as more primitive than the notion of reference (especially to the real world) for purposes such as a definition of coreference. We can use this insight to explain why substitutability does not hold in these examples — citing properties of vague referents, or of blended spaces — and get alternative criteria for these cases.

In the case of an actor and his role being interchangeably used for each other, for example, it is possible to consider complex actor/role entities that Fauconnier’s theory of mental spaces would predict to occur in a blended space – it then becomes possible to annotate these as equivalent descriptions even though they are not substitutable in the stricter sense. In the cases of sloppy reference discussed in the last subsection, the more complex account of reference to entities in the mental model helps to understand the ambiguity problem in terms of coercion (one description evokes several entities from the mental model), but does not lead to an obvious solution by itself.

Some researchers argue that the best solution to ambiguity problems is not to care at all, but leave contradictory judgements of annotators in the gold standard if both versions are plausible. This has actually been proposed by Strassel et al. (2004) for the ACE annotation effort, with the argument that the desire to always specify one single alternative as the right one leads to both increased cost and unnecessarily bloated manuals. This may be fine if the annotated corpus is just used as training or evaluation material for automatic classification or tagging, since it leads to evaluation scores that are somewhat closer to what people expect (i.e. fewer people will be discontent because their system has tagged what they think is the right solution, but the gold standard has been annotated differently). But if we wanted to take this at face value and say that we are really annotating the ambiguities in the corpus, we would be simply wrong. If two annotators agree, this is not a safe sign that a third annotator would not come up with a totally new solution, and if they disagree, we cannot be sure that one of the solutions isn't just marginal and would be annotated differently by the same annotator on a different day. While the original (and ambiguous) annotations are interesting as raw material, they fail to meet the predictability criterion almost as badly as a version annotated by only a single annotator would.

Poesio et al. (2006) go a step further and propose annotating ambiguities. In an experiment with 18 annotators annotating the same texts, they specifically encouraged annotators to annotate ambiguities. They found perfect or near-perfect agreement for 53% (79 out of 148) of the markables, and of the 67 markables that were implicitly ambiguous (i.e. more than one possibility chosen by more than two annotators), 38 were marked as ambiguous by at least one annotator (24 by at least two). While these results can be called quite encouraging, they also show that annotators are not really good at spotting ambiguities – in average, each coder that marked any items at all (three did not annotate any ambiguities) marked a little less than 9 markables. Arguably, better annotator training and the refinement of the ambiguous annotation in the revision process would improve this somewhat. But this kind of second-order ambiguity (between an ambiguous and an unambiguous annotation) means that annotators must be able to reliably distinguish ambiguities (due to sloppy reference) from misunderstandings or errors, a problem which becomes more acute when annotation is cast as a collaborative process to catch a maximum of ambiguities.[14] In the case of multiple mental spaces (such as in example 3), locally annotating ambiguities would also miss an important point since the ambiguities are indicative of the underlying structure. On the other hand, we almost certainly do not want to annotate *all* the structure that the theory of mental spaces predicts, since the predictions of this theory become unreliable in other aspects.

A sensible solution for this might be to try to minimize both annotated ambiguity and also the structure beyond local ambiguities that is annotated: In cases where a preferred mapping between sets of textual entities and discourse referents

[14] Poesio (p.c.) mentions that this will be done in an upcoming project called *AnaWiki*

(corresponding to a minimal model) can be constructed, as in examples 2 and 3, this should be annotated as unambiguous coreference, even though it violates substitutability; a light-weight theory such as the one sketched in this article helps distinguish these cases from real ambiguities even when substitutability does not hold. In cases which cannot be made unambiguous, such as the examples from subsection 4.3, annotation of ambiguities can then be used, as we hope, with greater reliability.

In most annotation projects, however, the overhead of explicitly annotating ambiguities will mean that a solution will be sought that results in an unambiguous annotation in all cases. This is still possible by explicitly stating preference rules such as those used by Chiarcos and Krasavina (2005). For example, their annotation guidelines include a rule to mark pronouns that are ambiguous between an idiomatic reading and an anaphoric reading should be marked as coreferent with their (potential) antecedent, but also marked as ambiguous. To extend this treatment to cases of referential ambiguity, and make one possible interpretation part of the official/formal gold standard annotation (thus ensuring theory-neutrality since we do not commit ourselves to annotate structures that only make sense within a given theory; information on ambiguities would be left as an informal part of the annotation to be exploited when revising the treebank or by researchers specifically interested in the ambiguities), it is necessary to provide meaningful criteria that allow for (i) a consistent identification of ambiguous cases and (ii) in the more complicated cases where the ambiguity is between several (otherwise plausible, but incompatible) coreference links, allow the selection of one of the alternatives. These criteria will have to be formulated using concepts from a 'lightweight theory' that can be used to analyse more complex cases and relates to full theories of reference where appropriate.

This kind of theory-neutral, theory-guided annotation can also provide a possible (partial) solution for manual bloat: because we essentially rely on existing theoretical frameworks, we only need the mapping from (finer) structures predicted by the theory to (coarser) equivalence relations in our coreference scheme, and not a full explanation of ad-hoc decisions (or possibly, partial generalizations of ad-hoc decisions).

6 Conclusion

After describing the benefits of creating large annotated corpora, we argued for the necessity for iterative improvement of the annotation and its documentation, backed by a light-weight theory that helps ensure the consistency of the annotation, to eventually obtain the highest quality possible. We also demonstrated the use of per-class agreement measures as a tool for post-hoc quantitative investigation of annotation problems. Using such an investigation, we could demonstrate that (plural-marked) groups of persons exhibit a much higher error rate than (singular) persons and organizations (which could be seen as a formally constituted group of persons). This is worth noting since, even when limiting coreference annotation to persons in an attempt to eliminate problematic cases, we would get similar problems for groups of persons as for events, since both are vague referents.

With an appropriate theory, in our case a combination of Karttunen's idea of discourse referents together with Smith and Brogaard's account of referential vagueness, we argue, one can make a principled argument that allows for consistent decisions in many cases. However, there are ambiguities that cannot be resolved due to several referents (vague or not) being plausible antecedents for an anaphoric description or pronoun. Poesio et al. (2003) propose an account based on their Justified Sloppiness Hypothesis that covers some of these cases. With a reformulation of this hypothesis into what we call the Generalized Sloppiness Hypothesis, we are able to cast the problem in terms of coercion, treating granularity shifts and structural polysemy similarly, and are able to explain some cases of ambiguity that the original version did not, and we hope that using the notion of dot objects introduced by coercion and granularity shifts will be generally helpful for discussing problems in coreference annotation.

Acknowledgements. I would like to thank Mareile Knees, Piklu Gupta, and Emily Jamison for helpful comments on various versions of this article, the two anonymous reviewers of this article for the detailed and very constructive criticism, as well as the editors of this issue, Massimo Poesio and Ron Artstein, for some very insightful suggestions. The referential annotation of TüBa-D/Z was carried out by Karin Naumann and Vera Möller (principal annotators) as well as Kathrin Eichler and Simone Müller (parallel annotation on selected parts). The author's work, as well as the annotation of TüBa-D/Z, has been supported by the Deutsche Forschungsgemeinschaft as part of the collaborative research centre "Linguistic Data Structures" (SFB 441).

References

Asher, N.: 2006, 'Things and Their Aspects'. *Philosophical Issues* **16**(1), 1–23.

Asher, N. and A. Lascarides: 2003, *Logics of Conversation*. Cambridge University Press.

Asher, N. and J. Pustejovsky: 2005, 'Word Meaning and Commonsense Metaphysics'. http://semanticsarchive.net/Archive/TgxMDNkM/.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal: 2006, 'The SALSA Corpus: a German corpus resource for lexical semantics'. In: *Proc. of LREC 2006*.

Cahill, A., M. McCarthy, J. van Genabith, and A. Way: 2002, 'Parsing with PCFGs and Automatic F-Structure Annotation'. In: *Proceedings of the Seventh International Conference on LFG*. CSLI Publications.

Carletta, J.: 1996, 'Assessing agreement on classification tasks: the kappa statistic'. *Computational Linguistics* **22**(2), 249–254.

Castaño, J., J. Zhang, and J. Pustejovsky: 2002, 'Anaphora Resolution in Biomedical Literature'. In: *International Symposium on Reference Resolution*.

Charniak, E. and M. Johnson: 2005, 'Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking'. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Chiarcos, C. and O. Krasavina: 2005, 'PoCoS – Potsdam Coreference Scheme'. Technical report, SFB 632 "Information structure: the linguistic means for structuring utterances, sentences and texts".

Cicchetti, D. V. and A. R. Feinstein: 1990, 'High agreement but low kappa: II. Resolving the paradoxes'. *J Clin Epidemiol.* **43**(6), 551–558.

Cohen, J.: 1960, 'A coefficient of agreement for nominal scales'. *Education and Psychological Measurement* **43**(6), 37–46.

Di Eugenio, B. and M. Glass: 2004, 'The Kappa statistic: a second look'. *Computational Linguistics* **30**(1), 95–101.

Dickinson, M. and W. D. Meurers: 2005, 'Prune Diseased Branches to Get Healthy Trees! How to Find Erroneous Local Trees in a Treebank and Why It Matters'. In: *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain.

Fauconnier, G.: 1984, *Espaces Mentaux*. Editions de Minuit.

Gardent, C. and H. Manuélian: 2005, 'Création d'un corpus annoté pour le traitement des déscriptions d éfinies'. *Traitement Automatique des Langues* **46**(1), 115–140.

Hinrichs, E., S. Kübler, and K. Naumann: 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., P. Robinson, J. Burger, and M. Vilain: 1997, 'Automating Coreference: The role of automated training data'. In: *Proc. of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Hobbs, J.: 1985, 'Granularity'. In: *Proc. IJCAI 1985*.

Hockenmaier, J. and M. Steedman: 2002, 'Acquiring Compact Lexicalized Grammars from a Cleaner Treebank'. In: *Proc. LREC 2002*.

Hoste, V. and W. Daelemans: 2004, 'Learning Dutch Coreference Resolution'. In: *Fifteenth Computational Linguistics in the Netherlands Meeting (CLIN 2004)*.

Hripcsak, G. and A. S. Rothschild: 2005, 'Agreement, the F-measure, and reliability in information retrieval'. *Journal of the American Medical Informatics Association* **12**, 296–298.

Karttunen, L.: 1976, 'Discourse Referents'. In: J. D. McCawley (ed.): *Syntax and Semantics 7: Notes from the Linguistic Underground.* Academic Press, pp. 363–385.

Kintsch, W. and T. van Dijk: 1978, 'Toward a model of text comprehension and production'. *Psychological Review* **85**, 363–394.

Knees, M.: 2006, 'The German Temporal Anaphor danach - Ambiguity in Interpretation and Annotation'. In: *ESSLLI 2006 workshop on Ambiguity and Anaphora.*

Link, G.: 1983, 'The logical analysis of plurals and mass terms: a lattice-theoretical approach'. In: R. Bäuerle, C. Schwarze, and A. von Stechow (eds.): *Meaning, Use and Interpretation of Language.* de Gruyter.

Luo, X., A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos: 2004, 'A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree'. In: *ACL 2004.*

Magerman, D. M.: 1995, 'Statistical Decision-tree Models for Parsing'. In: *ACL'1995.*

Mani, I.: 1998, 'A Theory of Granularity and its Application to Problems of Polysemy and Underspecification of Meaning'. In: A. G. Cohn, L. K. Schubert, and S. C. Shapiro (eds.): *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth Internatinal Conference (KR'98).* San Mateo, pp. 245–255, Morgan Kaufmann.

McCarthy, J. F. and W. G. Lehnert: 1995, 'Using Decision Trees for Coreference Resolution'. In: *IJCAI 1995.* pp. 1050–1055.

Meurers, W. D.: 2005, 'On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German'. *Lingua* **115**(11), 1619–1639.

Miyao, Y. and J. Tsujii: 2005, 'Probabilistic disambiguation models for wide-coverage HPSG parsing.'. In: *ACL 2005.*

MUC6: 1995, 'MUC-6 Coreference Task Definition'. DARPA Information Technology Office Tipster Text Program.

Passonneau, R.: 1997, 'Applying reliability metrics to co-reference annotation'. Technical Report CUCS-025-03, Columbia University.

Poesio, M.: 2000, 'The GNOME Annotation Scheme Manual'. Technical report, University of Edinburgh, HCRC and Informatics. http://www.hcrc.ed.ac.uk/˜gnome.

Poesio, M.: 2004, 'The MATE/GNOME Scheme for Anaphoric Annotation, Revisited'. In: *Proc. of SIGDIAL'04*. Boston.

Poesio, M. and R. Artstein: 2005, 'Annotating (Anaphoric) Ambiguity'. In: *Corpus Linguistics 2005*. Birmingham.

Poesio, M. and U. Reyle: 2001, 'Underspecification in Anaphoric Reference'. In: *Fourth International Workshop on Computational Semantics (IWCS-4)*.

Poesio, M., U. Reyle, and R. Stevenson: 2003, 'Justified Sloppiness in Anaphoric Reference'. In: H.Bunt and R.Muskens (eds.): *Computing Meaning 3*. Kluwer. to appear.

Poesio, M., P. Sturt, R. Artstein, and R. Filik: 2006, 'Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence'. *Discource Processes* **42**(2), 152–175.

Reitsma, F. and T. Bittner: 2003, 'Process, Hierarchy and Scale'. In: W. Kuhn, M. Worboys, and S. Timpf (eds.): *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science (COSIT'03)*.

Setzer, A. and R. Gaizauskas: 2001, 'A Pilot Study on Annotating Temporal Relations in Text'. In: *ACL 2001 Workshop on Temporal and Spatial Information Processing*.

Smith, B. and B. Brogaard: 2001, 'A Unified Theory of Truth and Reference'. *Logique et Analyse* **43**(169-170), 49–93.

Strassel, S., C. Walker, and A. Mitchell: 2004, 'Annotation Consistency Study'. Slides found at http://projects.ldc.upenn.edu/ace/workshops/Feb2004.html.

Uryupina, O.: 2006, 'Coreference Resolution with and without Linguistic Knowledge'. In: *Proceedings of LREC 2006*.

van Deemter, K. and R. Kibble: 2000, 'On Coreferring: Coreference in MUC and Related Annotation Schemes'. *Computational Linguistics* **26**(4), 629–637.

van Rijsbergen, C. J. K.: 1979, *Information Retrieval*. Butterworths.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman: 1995, 'A model-theoretic coreference scoring scheme'. In: *Proceedings of the 6th Message Understanding Conference*.

Zaenen, A., J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitana, M. C. O'Connor, and T. Wasow: 2004, 'Animacy Encoding in English: why and how'. In: *ACL 2004 Workshop on Discourse Annotation*.