# Annotating genericity: a survey, a scheme, and a corpus

**Annemarie Friedrich**[1] **Alexis Palmer**[2] **Melissa Peate Sørensen**[1] **Manfred Pinkal**[1]

[1]Department of Computational Linguistics, Universität des Saarlandes, Germany
{afried,melissap,pinkal}@coli.uni-saarland.de

[2]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany
alexis.palmer@ims.uni-stuttgart.de

## Abstract

*Generics* are linguistic expressions that make statements about or refer to kinds, or that report regularities of events. *Non-generic* expressions make statements about particular individuals or specific episodes. Generics are treated extensively in semantic theory (Krifka et al., 1995). In practice, it is often hard to decide whether a referring expression is *generic* or *non-generic*, and to date there is no data set which is both large and satisfactorily annotated. Such a data set would be valuable for creating automatic systems for identifying generic expressions, in turn facilitating knowledge extraction from natural language text. In this paper we provide the next steps for such an annotation endeavor. Our contributions are: (1) we survey the most important previous projects annotating genericity, focusing on resources for English; (2) with a new agreement study we identify problems in the annotation scheme of the largest currently-available resource (ACE-2005); and (3) we introduce a linguistically-motivated annotation scheme for marking both clauses and their subjects with regard to their genericity. (4) We present a corpus of MASC (Ide et al., 2010) and Wikipedia texts annotated according to our scheme, achieving substantial agreement.

## 1 Introduction

This paper addresses the question of distinguishing clauses or noun phrases (NPs) that convey information about particular entities or situations, as in example (1a), from those which convey general information about kinds, see example (1b).

*(1)* *(a) Simba is in danger. (**non-generic**)*
*(b) Lions live for 10–14 years. (**generic**)*

Making this distinction is important for NLP tasks that aim to disentangle information about particular events or entities from general information about classes, kinds, or particular individuals, such as question answering or knowledge base population. Our present work targets the current lack of a large and satisfactorily-annotated data set for genericity, which is a prerequisite for research aiming to automatically identify these linguistic phenomena.

Krifka et al. (1995) report the central results in semantic theory on *genericity*. Several phenomena have been studied within this research field: one is *reference to a kind*, which is a NP-level property. The form of the NP itself (definite, indefinite, ...) is not sufficient to make this distinction (Carlson, 1977; Chierchia, 1998); the interpretation of the NP depends on the clause in which it appears, see (2).

*(2)* *The lion is a predatory cat. (**kind-referring**)*
*The lion escaped from the zoo. (**non-generic**)*

*Characterizing sentences* are another phenomenon studied under the heading of genericity. They may be *lexically characterizing*, as in (3a) and (3b), or *habitual* as in (3c) and (3d). Habitual sentences describe regularly occurring episodes rather than specific ones. Characterizing sentences as in (3) may relate to a kind (*lions*), or to a particular individual (*John*).

*(3)* *(a) Lions have manes.*
*(b) John is tall.*
*(c) Lions eat meat.*
*(d) John drives to work.*

*Proceedings of LAW IX - The 9th Linguistic Annotation Workshop*, pages 21–30,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

Statements about kinds, such as example (3a), are not rendered false by the existence of counterexamples. If we encountered a vegetarian lion, it would still be true that a *typical* lion eats meat. Such sentences have been analyzed as referring to a kind instead of a set of entities (Carlson, 1977), or as containing a 'generic' quantifier (Krifka et al., 1995). Similarly, habitual sentences such as (3d) are not rendered false by exceptions.

As the linguistic manifestations of both generic and non-generic clauses (and NPs) are quite diverse, automatic discrimination between generic and non-generic information is a highly-challenging task, and annotated resources are necessary for making progress. Existing corpora for genericity focus on different aspects of genericity or related phenomena.

In this paper we provide a comprehensive survey of existing resources for computational treatment of genericity (Section 2). Section 3 presents an agreement study for ACE-2005, the largest annotation project regarding genericity of NPs to date, highlighting problems in their annotation scheme.

In Section 4, we introduce a linguistically motivated annotation scheme for marking genericity. We focus both on whether a clause makes a characterizing statement about a kind and whether its subject refers to a kind, eliminating some of the uncertainties in some previously-proposed schemes. Our scheme does not address whether a clause is habitual or not, leaving this question to future work. We apply our scheme to several sections of the Manually Annotated SubCorpus (MASC) of the Open American National Corpus (Ide et al., 2010) and to Wikipedia texts, mostly reaching substantial agreement.

## 2 Survey: annotating genericity in English

Existing resources treat both NP- (Section 2.1) and clause-level (Section 2.2) phenomena related to genericity. For each approach, we explain the annotation scheme, discuss its relation to theoretical concepts, and describe the data labeled. Table 1 gives a summary.

### 2.1 NP-level annotations

Section 2.1.1 describes corpora from the Automatic Content Extraction (ACE) program (Doddington et al., 2004); other NP-level approaches are described in Section 2.1.2.

#### 2.1.1 ACE entity class annotations

The research objective of the ACE program (1999-2008) was the detection and characterization of entities, relations and events in natural text (Linguistic Data Consortium, 2000). All entity mentions receive an *entity class* label indicating their genericity status. Of the corpora described here, the ACE corpora have been the most widely used for recent research on automatically identifying generic NPs (Reiter and Frank, 2010). The annotation guidelines developed over time; we describe both the initial guidelines of ACE-2 and those from ACE-2005.

The **ACE-2 corpus** (Mitchell et al., 2003) includes 40106 annotated entity mentions in 520 newswire and broadcast documents. The annotation guidelines give no formal definition of genericity; annotators are asked to determine whether each entity refers to "any member of the set in question" (**generic**) or rather "some particular, identifiable member of that set" (**specific/non-generic**).[1] This leads to a mix of constructions being marked as generic: types of entities (*Good students do all the reading*), generalizations across a set of entities (*Purple houses are really ugly*), hypothetical entities (*If a person steps over the line,...*) and negated mentions (*I saw no one*). Suggested attributes of entities are marked as generic (*John seems to be a nice person*), but a 'positive assertion test' leads to marking both NPs (*Joe* and *a nice guy*) as specific in examples like (*Joe is a nice guy*). Neither of these two cases (*be a nice person / be a nice guy*) is in fact an entity mention; they are rather predicative uses.

The guidelines for genericity were redefined for annotation of the **ACE-2005 Multilingual Training Corpus** (Walker et al., 2006), which contains news, broadcast news, broadcast conversation, forum and weblog texts as well as transcribed conversational telephone speech. In contrast to ACE-2, the ACE-2005 annotation manual[2] clearly defines mentions as kind-referring or not, using the labels GEN (generic)

[1] See "Entity Detection Tracking and Metonymy Annotation Guidelines, Version 2.5.1", available from LDC: `https://catalog.ldc.upenn.edu/docs/LDC2003T11/`

[2] See "ACE English Annotation Guidelines for Entities, Version 5.6.6" (available from LDC) or 2008's version 6.6.

| Corpus | Level | Scheme | Amount |
|---|---|---|---|
| ACE-2 | NP | generic, specific | 40K entity mentions |
| ACE-2005 | NP | GEN, SPC, USP, NEG | 40K entity mentions |
| ECB+ | NP | GEN, non-GEN | 12.5K entity mentions |
| GNOME | NP | generic-yes, generic-no | 900 clauses |
| Herbelot & Copestake | NP | One, Some, Most, All, Quant | 300 subject mentions |
| CFD | NP | Generic_Kind, Generic_Individual | 3422 NPs (131 generic) |
| Mathew & Katz | clause | habitual, episodic | 1052 sentences |
| Louis & Nenkova | clause | general, specific | 894 sentences |
| MASC | NP, clause | GEN_gen, NON-GEN_gen, NON-GEN_non-gen | 20K clauses |
| WikiGenerics | NP, clause | | 10K clauses |

Table 1: **Survey of genericity-annotated corpora** *for English*, including our new corpus.

and SPC (specific/non-generic) respectively.

The new guidelines also introduce two additional entity class labels for non-attributive mentions. Negatively quantified entities that refer to the empty set of the kind mentioned (*There are no confirmed suspects yet*) receive the label NEG. The label USP (underspecified) is used for non-generic nonspecific reference, these cases include quantified NPs in modal, future, conditional, hypothetical, negated, uncertain or question contexts. USP also covers 'truly ambiguous cases' that have both generic and non-generic readings (*The economic boom is providing new opportunities for women in New Delhi*), and cases where the author mentions an entity whose identity would be 'difficult to locate' (*Officials reported ...*). In our opinion, the latter interferes with the definition of SPC as marking cases where the entity referred to is a particular object in the real world, even if the author does not know its identity (*At least four people were injured*). The breadth of the USP category causes problems with consistency of application (see Section 3).

The ACE annotation scheme has also been applied in the **Newsreader** project.[3] The **ECB+ corpus** (Cybulska and Vossen, 2014) is an extension of EventCorefBank (ECB), a corpus of news articles marked with event coreference information (Bejan and Harabagiu, 2010). ECB+ annotates entity mentions according to ACE-2005, but collapses the three non-GEN labels into a single category. Roughly 12500 event participant mentions are annotated, some doubly and some singly. Agreement statistics for genericity are not reported.

### 2.1.2 Other corpora annotated at the NP-level

The resources surveyed here apply carefully-defined notions of genericity but are too small to be feasible machine learning training data.

The question of whether an NP is generic or not arises in the research context of coreference resolution. Some approaches mark coreference only for non-generic mentions (Hovy et al., 2006; Hinrichs et al., 2004); others include generic mentions (Poesio, 2004), or take care not to mix coreference chains between generic and non-generic mentions (Björkenstam and Byström, 2012). Björkelund et al. (2014) mark genericity in a corpus of German with both coreference and information-status annotations. Nedoluzhko (2013) survey the treatment of genericity phenomena within coreference resolution research; they provide a complete overview. In short, they argue that a consistent definition of genericity is lacking and report on their annotation scheme for Czech as applied to the Prague Dependency TreeBank (Böhmová et al., 2003).

The **GNOME corpus** (Poesio, 2004) is a coreference corpus with genericity annotations; NPs are marked with the attributes `generic-yes` or `generic-no`. Poesio et al. report that their annotators found it hard to decide how to mark references to substances (*A table made of wood*) and quantified NPs. Similar to our experience, they found it helpful to have annotators first try to identify generic sentences, and then determine this attribute of the NP. They report an agreement of $\kappa = 0.82$ on their corpus, which consists of 900 finite clauses from descriptions of museum objects, pharmaceutical leaflets and dialogues.

[3] www.newsreader-project.eu

Coming from a formal semantic perspective, Herbelot and Copestake (2010) and Herbelot and Copestake (2011) describe an approach to treating **ambiguously quantified NPs**. This annotation effort aims to produce resources for the task of determining the extent to which the semantic properties ascribed to a given NP in context apply to the members of that class. For example, the statement *Cats are mammals* describes a property of *all* cats, where *Cats have four legs* is true only for most cats. The scheme, which includes the labels ONE, SOME, MOST, ALL and QUANT (for explicitly quantified NPs), is applied to 300 subject-verb-object triples from sentences randomly extracted from Wikipedia. Annotators are shown the sentence and the triple. $\kappa$ ranges from 0.88 and 0.81 for QUANT and ONE to values between 0.44 and 0.51 for the other classes.

Bhatia et al. (2014b) present an annotation scheme for **Communicative Functions of Definiteness**, intended to cover the many semantic and pragmatic functions conveyed by choices regarding definiteness across languages of the world. The scheme has been applied to 3422 English NPs contained in texts from four genres. Their typology includes two categories relevant to our survey: GENERIC_KIND_LEVEL applies to utterances predicating over an entire class, like *Dinosaurs are extinct*. GENERIC_INDIVIDUAL_LEVEL is for predications applying to the individual members of a class or kind, such as *Cats have fur*. Across 1202 annotated NPs for an inter-annotator agreement study, the two annotators used the GENERIC_INDIVIDUAL_LEVEL label 45 times and 30 times, respectively, with agreement in 29 cases. Neither used the GENERIC_KIND_LEVEL. The entire corpus contains just 131 NPs labeled with GENERIC_INDIVIDUAL_LEVEL and none with GENERIC_KIND_LEVEL (Bhatia et al., 2014a).

The question of genericity has also been addressed in cognitive science (Prasada, 2000). Gelman and Tardif (1998) study the usage of generic NPs cross-linguistically for English and Chinese in child-directed speech. They annotate kind-referring NPs as generic. They report agreement as the fraction of items on which the annotators agreed at over 99%, but given that their data set has fewer than 1% generic NPs, this statistic does not allow us to estimate how well annotators agreed.

### 2.2 Clause-level annotations

The two resources described in this section are the only we know of which mark phenomena related to genericity on clauses of text.

**Annotating habituality.** Mathew and Katz (2009) conduct a study on automatically distinguishing *habitual* from *episodic* sentences. Habitual sentences are taken to be sentences whose main verb is lexically dynamic, but which do not refer to particular events (see for example (3)), and may have generic or non-generic subjects. Their singly-annotated data set, from which they excluded verb types with skewed class distributions, comprises 1052 examples covering 57 verb stems. Their data set is not publicly available.

**General vs. specific sentences.** Louis and Nenkova (2011) describe a method for automatic classification of sentences as *general* or *specific*. *General* sentences are loosely defined as those which make "broad statements about a topic," while *specific* sentences convey more detailed information. This distinction is not immediately related to the phenomena treated as generics in the literature. Kind-referring subjects can occur in both *general* (4a) and *specific* (4b) sentences; *general* sentences can also have non-kind-referring subjects (4c).

*(4)* *(a) Climatologists and policy makers, ..., need to ponder such complexities... (**general**)*
*(b) Solid silicon compounds are already familiar – as rocks, glass, ... (**specific**)*
*(c) A handful of serious attempts have been made to eliminate ... diseases. (**general**)*

## 3 ACE-2005: an agreement study

In this section we investigate some problems with the ACE annotation scheme via a study of annotator agreement. The data was first labeled by two annotators independently, then adjudicated by a senior annotator. To our knowledge, agreement numbers on this task have not been published to date. In order to assess both the quality of the data and the difficulty of the task, we compute inter-annotator agreement as follows. Using the 533 documents from the adjudicated data set that were marked by two annotators in the first step, we compute Cohen's $\kappa$ (Cohen, 1960) for entity class annotations over the four labels SPC, GEN, USP and NEG.

Intuitions about NP genericity are most reliable for subject position as other argument positions involve additional difficulties (Link, 1995). To get a better sense of the difficulty of annotating subjects compared to that for other argument positions, we compute agreement over mentions whose (manually marked) head is the grammatical subject of some other node in a dependency graph (including any dependency type containing `subj`). We obtain dependency graphs using the Stanford parser (Klein and Manning, 2002).

An additional complication in entity mention annotation is determining the mention span. Because spans are not pre-marked in the ACE corpora but identified independently by each annotator, we compute $\kappa$ only over all exactly-matching entity mention spans for the two annotators. For all mentions, annotators mark about 90% of spans marked by the other annotator. For subject mentions, this number is even higher, at about 95%. The spans of the remaining mentions overlap for the two annotators. We exclude them from this study as we cannot be sure that the two mention spans refer to the same entity.

**Discussion.** Table 2 shows the confusion matrices of labels for the all-mentions-case and the subjects-only case. In both cases, confusion between SPC and GEN is acceptable, but confusion between USP and both SPC and GEN is rather high. For example, in the case of subjects, annotator 1 tags 652 mentions as GEN that annotator 2 marks USP, but the two of them only agree on 597 mentions to be GEN. Although it may be useful to create a separate category for unclear or underspecified cases, the definition of USP is not yet clear-cut and compounded with lack of *specificity*, which refers to whether the speaker presumably knows the referent's identity or not. Even if the identity of a referent may be 'difficult to locate' (as in *Officials reported...*). The clause certainly does not make a statement about the *kind* 'official'; instead, it expresses an existential statement (*There are officials who reported...*). The definition of SPC states that the reader does not necessarily have to know the identity of the entity, possibly making the distinction hard for annotators.

Another difficult case are noun modifiers in compounds (e.g. *a subway system*); these are marked as GEN in the corpus. Using the automatic parses, we find that 9.5% of all mentions marked GEN in the adjudicated corpus are one-token mentions modifying another noun via an *nn* dependency relation. Genericity as reference to kinds is a discourse phenomenon and thus defined as an attribute of referring expressions. Because nominal modifiers do *not* introduce discourse referents, they should not be treated on the genericity annotation layer.

| | | *annotator 2* | | | |
|---|---|---|---|---|---|
| **all mentions** | | SPC | USP | GEN | NEG |
| *annotator 1* | SPC | 28168 | 1575 | 684 | 3 |
| | USP | 1142 | 1954 | 963 | 2 |
| | GEN | 757 | 1261 | 1707 | 10 |
| | NEG | 8 | 5 | 7 | 71 |

| | | *annotator 2* | | | |
|---|---|---|---|---|---|
| **subjects only** | | SPC | USP | GEN | NEG |
| *annotator 1* | SPC | 9830 | 830 | 234 | 1 |
| | USP | 634 | 1091 | 476 | 1 |
| | GEN | 272 | 652 | 597 | 4 |
| | NEG | 4 | 1 | 2 | 46 |

Table 2: **Confusion matrices** of entity class tags for ACE 2005 for mentions where annotators agree on spans.

The data shows moderate agreement for the first two passes of entity class annotation ($\kappa = 0.53$ for all mentions and $\kappa = 0.50$ for subject mentions). Note that $\kappa$ scores are not directly comparable across different annotation projects (see also Section 5), we give the above scores for the sake of completeness. Observed and expected agreement are 0.83 and 0.65 for the all-mentions case and 0.79 and 0.58 for subject mentions. This indicates that the all-mentions case may contain some trivial cases, one of which is the case of nominal modifiers described above.

In summary, the ACE scheme problematically fails to treat subject NPs differently from NPs in other syntactic positions, and 'fuzzy' points in the guidelines, particularly concerning the USP label, contribute to disagreements between annotators.

## 4 Annotating genericity as reference to kinds on NP- and clause-level

We next present an annotation scheme for marking both clauses and their subject NPs with regard to whether they are generic. Our scheme is primarily motivated by the contributions of clauses to the discourse (Friedrich and Palmer, 2014): do they report on a particular event or state, or do they report on some regularity? These different types of clauses have different entailment properties, and differ in how they contribute to the temporal structure of the discourse. In this work, we focus on separating generic clauses from other types of clauses. We approach the problem from a linguistic perspective rather than focusing on any particular content extraction task, arguing that any generally applicable annotation scheme must be based on solid theoretical foundations. We believe our annotation scheme is a step toward solving the problems of marking genericity in natural text. We apply our annotation scheme to two text corpora[4], reaching substantial agreement on Wikipedia texts.

### 4.1 Annotation scheme

The definition of our annotation scheme is guided by the following questions: (a) does a clause's subject refer to a kind rather than a particular individual; (b) if so, does the clause make a characterizing statement about the kind or its members, or does it report a particular episode related to the kind?

**Task NP: genericity of subject.** In this step, annotators decide whether the subject of the clause refers to a kind (**generic**) or to a particular individual (**non-generic**) as in (5d). In English, definite singular NPs (5a) or bare plural NPs (5b) can reference kinds. Indefinite singular NPs (5c) can refer to arbitrary members of a kind; these are also marked **generic**.

***(5)*** *(a) <u>The lion</u> is a predatory cat. (**generic**)*
*(b) <u>Lions</u> have manes. (**generic**)*
*(c) <u>A lion</u> may eat up to 30kg in one sitting. (**generic**)*
*(d) <u>Simba</u> the lion flees into exile. (**non-generic**)*

The label **non-generic** also includes cases of non-specific reference if the reader can infer that the clause makes a statement about some particular individual (or group of individuals), even if the identity is unknown, as (6a). This is precisely where the ACE guidelines are somewhat unclear, mixing annotation of genericity and specificity. We aim to convey and mark this difference clearly. In (6b), the determiner 'some' could be added without changing the meaning significantly, showing that the bare plural here is existential, not generic (Carlson, 1977).

***(6)*** *(a) <u>A lion</u> must have eaten the rabbit. (**non-specific**, **non-generic**)*
*(b) <u>Lions</u> are in this cage. (**non-generic**)*
*(c) <u>Dinosaurs</u> are extinct. (**generic**)*

**Task Cl: genericity of clause.** We define **generic** clauses as making characterizing statements about kinds. This includes both clauses predicating something directly of the 'kind individual' itself (6c) and clauses that predicate something of the members of a kind, such as (5b) and (5c). According to our definition, **generic** sentences may be lexically characterizing, as in (5a) or (5b), or they may describe something that members of the kind do regularly, as in (5c). The latter type of sentences are called *habituals*. The subject of a **generic** clause must necessarily be **generic**. In addition, episodic events, classified as **non-generic** clauses, can have a **generic** NP as their subject, as in example (7). Note that we mark any clause about particular individuals as **non-generic**, including habituals making a statement about particular individuals (8). The question of whether a clause with a **non-generic** subject is habitual or not is another interesting related question, but for the moment, we leave this to future work and concentrate on the distinction of whether a clause relates to kinds.

***(7)*** *In September 2013 <u>the blobfish</u> was voted the "World's Ugliest Animal". (**generic** subject, **non-generic** clause)*

***(8)*** *John cycles to work. (**non-generic**)*

**Task Cl+NP.** Using the information from Tasks NP and Cl, we automatically derive a combination label from the following set for each *clause*:

- **GEN_gen**: a **generic** clause, subject is **generic** by definition;
- **NON-GEN_non-gen**: a **non-generic** clause with a **non-generic** subject;
- or **NON-GEN_gen**: an episodic (**non-generic**) clause with a **generic** subject, see example (7).

The combination **GEN_non-gen** is not possible, by definition.

[4] The annotated corpora are freely available from `http://sitent.coli.uni-saarland.de`

|  | # documents | # clauses | Task NP | Task Cl | Task Cl+NP | % generic |
|---|---|---|---|---|---|---|
| botany | 6 | 592 | 0.68 | 0.70 | 0.69 | 77.8 |
| games | 5 | 567 | 0.61 | 0.63 | 0.59 | 77.4 |
| animals | 13 | 1924 | 0.66 | 0.70 | 0.67 | 65.6 |
| music | 12 | 861 | 0.76 | 0.75 | 0.74 | 61.3 |
| medicine | 7 | 561 | 0.72 | 0.78 | 0.73 | 59.8 |
| science | 8 | 711 | 0.62 | 0.66 | 0.60 | 47.0 |
| sports | 8 | 1242 | 0.70 | 0.72 | 0.67 | 43.1 |
| politics | 16 | 1466 | 0.62 | 0.65 | 0.61 | 40.9 |
| ethnic groups | 8 | 582 | 0.57 | 0.60 | 0.57 | 40.0 |
| religion | 8 | 622 | 0.57 | 0.62 | 0.58 | 35.7 |
| crime | 4 | 588 | 0.50 | 0.60 | 0.52 | 26.3 |
| biographies | 7 | 563 | 0.63 | 0.69 | 0.63 | 8.9 |
| all | 102 | 10279 | 0.69 | 0.72 | 0.68 | 50.1 |

Table 3: **IAA on WikiGenerics**. **Fleiss'** $\kappa$ for three annotators that marked the entire data set. % generic = percentage of clauses marked as generic in Task Cl according to the majority vote gold standard.

### 4.2 Corpus data: MASC/WikiGenerics

We apply the annotation scheme explained above to two corpora comprising texts of a wide range of genres and domains. We annotate several sections of the Manually Annotated SubCorpus (MASC) of the Open American National Corpus (Ide et al., 2010). In addition, we collect 102 texts from Wikipedia (**WikiGenerics** corpus) from a variety of categories (see Table 3). Our aim is to create a corpus that is balanced in the sense that it contains many generic and non-generic sentences, and also many different varieties of generic sentences. The corpus contains (among others) sentences about animals (9a), rule-like knowledge about sports and games (9b), and clauses describing abstract concepts (9c).

***(9)*** *(a) Blobfish are typically shorter than 30 cm.*
*(b) The offensive team must line up in a legal formation before they can snap the ball.*
*(c) A dictatorship is a type of authoritarianism.*

Note that we mark complete texts: the genericity of some sentences clearly depends on their context. For example, (9b) is **generic** as the text describes the rules of a game rather than a specific instance of the game.

We use the discourse parser SPADE (Soricut and Marcu, 2003) to segment the first 70 sentences of each Wikipedia article into clauses, which are the basis for annotation. Subjects are not pre-marked and do not necessarily have to have their mention spans in the same segment, as illustrated in (10).

***(10)*** *(a) Blobfish look funny (**GEN_gen**)*
*(b) and were voted the most ugly animal. (**NON-GEN_gen**)*

Annotators were allowed to skip clauses that do not contain a finite verb, which constitute about 5% of all pre-marked clauses. These clauses are mostly headlines consisting only of an NP.

### 4.3 Inter-annotator agreement

Our aim is to create a gold standard via majority voting. Annotators were given a written manual and a short training on documents not included in the corpus. The WikiGenerics corpus was marked completely by three paid annotators (students of computational linguistics), and agreement is given in Table 3 in terms of Fleiss' $\kappa$ (Fleiss, 1971). We observe substantial agreement in almost all categories, and moderate agreement in only three categories: games, ethnic groups and organized crime. The categories ethnic groups and organized crime were especially hard to annotate because they contain many cases where it is not clear whether a mention refers to a very large particular group or whether this group rather counts as reference to a kind, as in (11).

***(11)*** *The Bari also known as the Karo ethnic groups in South Sudan occupy the Savanna lands of the White Nile Valley.*

For MASC, two annotators mark each section; we report agreement as Cohen's $\kappa$ for these two annotators in Table 4. Then, a third annotator marks all

| section | # clauses | Task NP (subject) | Task Cl (clause) | Task Cl+NP (clause) | % generics |
|---|---|---|---|---|---|
| essays‡ | 1590 | 0.55 | 0.56 | 0.54 | 27.9 |
| travel† | 1922 | 0.38 | 0.45 | 0.41 | 19.0 |
| letters† | 1944 | 0.33 | 0.41 | 0.40 | 14.2 |
| journal† | 1927 | 0.42 | 0.52 | 0.48 | 13.0 |
| jokes† | 3376 | 0.56 | 0.63 | 0.58 | 11.6 |
| blog† | 2723 | 0.09 | 0.13 | 0.14 | 10.4 |
| news‡ | 2557 | 0.25 | 0.33 | 0.29 | 3.4 |
| fiction†* | 4124 | 0.50 | 0.59 | 0.54 | 2.5 |

Table 4: **IAA for MASC**. The sections were marked by different pairings of annotators: †Cohen's $\kappa$ for 2 annotators; ‡Fleiss' $\kappa$ for 3 annotators. *fiction: agreement for 70% of data that was marked by the same two annotators. % generic = percentage of clauses marked as generic in Task Cl according to the majority vote gold standard.

clauses on which the two annotators of the first step disagreed, without seeing the annotations of the first step. Hence, this does not constitute an adjudication step. Two sections, essays and news, were marked completely by three annotators. Five paid annotators, all students of computational linguistics, participated in the annotation of MASC. The various MASC sections show a larger variation both in the percentage of generic clauses and in the agreement numbers. News and fiction contain almost no generics, while essays, travel, and letters contain notable numbers. Agreement on the blog section is surprisingly low. One annotator rarely used the category **generic** here, while the other annotator did. Manual inspection showed that this section contains many intrinsically ambigous instances of 'you' and 'one'. The third annotator agrees well with the annotator who marked more clauses as generic.

**Discussion.** In general, $\kappa$ numbers are difficult to compare, as the expected agreement depends on the distribution of labels (Di Eugenio and Glass, 2004). If the distribution is skewed, the expected agreement is high and it is thus harder to reach a high $\kappa$ score. We give the percentage of clauses labeled as generic in Task Cl. A small percentage but a relatively high $\kappa$ score (as in the jokes section) means that in this category, it was apparently easier for the annotators to agree. For example, in the fiction genre, there are very few generics, but a high agreement was reached nonetheless. In the narratives of this subcorpus, the generics apparently 'stand out' clearly.

In this study, substantial agreement was reached on Wikipedia texts using our annotation scheme. The lower agreement reached on some MASC sections indicates that the annotation task is harder for some text types, and this difficulty is only partially explained by the skewedness of the label distribution: some genres simply contain more borderline cases than others.

## 5 Discussion and future work

We have proposed an annotation scheme for labeling clauses with regard to whether they make a characterizing statement about kinds, and NPs with regard to whether they refer to kinds or not. Our scheme aims at a linguistically motivated annotation in order to advance our understanding of generics and to see to what extent existing linguistic theories can be applied to natural text of various genres and domains.

Across all of the surveyed annotation studies and also in our own experience, agreement on the task of annotating genericity was moderate to substantial, however, $\kappa$-scores need to be interpreted in relation to the distribution of labels and are not directly comparable across different annotation projects. Annotating genericity is not an easy task even for trained annotators, as there are many borderline cases, which occur frequently in some texts and very infrequently in others. As future work, we want to investigate whether it is possible to reliably label such 'underspecified' cases, redefining ACE's USP class in a way that disentangles the annotation of genericity and specificity.

The present survey focuses on resources in English, and our new annotation scheme has only been worked out for English. We plan to extend the annotation scheme and corpus to other languages including German and Chinese.

**Acknowledgments.** We thank Andrea Horbach for helpful comments, and our annotators Fernando Ardente, Christine Bocionek, Ambika Kirkland, Ruth Kühn and Kleo Mavridou. This research was supported in part by the MMCI Cluster of Excellence, and the first author is supported by an IBM PhD Fellowship.

## References

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of ACL*.

Archna Bhatia, Chu-Cheng Lin, Nathan Schneider, Yulia Tsvetkov, Fatima Talib Al-Raisi, Laleh Roostapour, Jordan Bender, Abhimanu Kumar, Lori Levin, Mandy Simons, et al. 2014a. Automatic classification of communicative functions of definiteness. *Proceedings of COLING*, pages 1059–1070.

Archna Bhatia, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer, and Jordan Bender. 2014b. A unified annotation scheme for the semantic/pragmatic components of definiteness. In *Proceedings of LREC*.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Kristina Nilsson Björkenstam and Emil Byström. 2012. SUC-CORE: SUC 2.0 Annotated with NP Coreference. *Proceedings of SLTC*.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Gregory Norman Carlson. 1977. *Reference to kinds in English*. Ph.D. thesis.

Gennaro Chierchia. 1998. Reference to kinds across language. *Natural language semantics*, 6(4):339–405.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pages 37–46.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of LREC*.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of LREC*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII)*, page 149.

Susan A Gelman and Twila Tardif. 1998. A crosslinguistic comparison of generic noun phrases in English and Mandarin. *Cognition*, 66(3):215–248.

Aurelie Herbelot and Ann Copestake. 2010. Annotating underquantification. In *Proceedings of the Fourth Linguistic Annotation Workshop*.

Aurelie Herbelot and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the International Conference on Computational Semantics*.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. 2004. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, pages 51–62.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of HLT-NAACL: Short Papers*, pages 57–60.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of ACL: Short papers*, pages 68–73.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10.

Manfred Krifka, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An Introduction. *The Generic Book*, pages 1–124.

Linguistic Data Consortium. 2000. Entity Detection and Tracking - Phase 1, ACE Pilot Study Task Definition. https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications.

Godehard Link. 1995. Generic information and dependent generics. *The Generic Book*, pages 358–382.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*.

Thomas A. Mathew and E. Graham Katz. 2009. Supervised categorization of habitual and episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium*, Bloomington, Indiana: Indiana University.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. 2003. ACE-2 Version 1.0 LDC2003T11. Philadelphia: Linguistic Data Consortium.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 72–79. Association for Computational Linguistics.

Sandeep Prasada. 2000. Acquiring generic knowledge. *Trends in cognitive sciences*, 4(2):66–72.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of ACL*, pages 40–49, Uppsala, Sweden.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL-HLT*, pages 149–156.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus LDC2006T06. Philadelphia: Linguistic Data Consortium.