Annotating Genericity: How Do Humans Decide?
(A Case Study in Ontology Extraction)

Aurelie Herbelot & Ann Copestake
Computer Laboratory
University of Cambridge

1. Introduction

In computational linguistics, the task of ontology extraction deals with acquiring factual statements from natural language text. Those statements are traditionally added to knowledge bases (so-called 'ontologies') in the form of relationships linking one or several concepts together. The relations can either take the form of general statements such as 'whale – is a – mammal' or of more anecdotal information such as 'whale – escape from – zoo'.

The research presented here is part of a project aimed at the construction of a tool able to summarise – or rather sketch – the general ideas of a domain using such an ontological representation. The final software is expected to:

- transform a domain-specific corpus into a light ontology consisting of generic triples of the type 'A does B' (by 'generic', we mean that we are not interested in anecdotal or exemplary information, but in general statements about the world – the 'beliefs' of the domain).
- conflate related triples into topical clusters, to give the main 'themes' of the discourse
- find evidence for or against individual statements using the web and return them as basic 'critiques' of the ontology
- allow the user to query the ontology in their own words (a query on 'houses' should also return statements on 'homes'.)

This paper deals with the first point, the extraction of generic statements from corpora. There is no proposal in the current ontological research for distinguishing generic from specific statements at extraction stage. If this does not generally affect systems dealing with strictly-defined relations (like those acquiring chemical reactions, Nobel Prize winners or company president successions), it does have an impact on the performance of systems extracting more general relations (like taxonomy or meronymy) and those attempting to give some ontological representation of a given text.

Table 1 a list of relationships that might be extracted from the following paragraph (taken from the Wikipedia article on grey whales):

The grey whale feeds mainly on benthic crustaceans which it eats by turning on its side (usually the right) and scooping up the sediments (usually on the right) from the sea floor. It is classified as a baleen whale and has a baleen, or whalebone, which acts like a sieve to capture amphipods taken in along with sand, water and other material. Mostly, the animal feeds in the northern waters during the summer; and opportunistically feeds during its migration trip, and mainly lives off its extensive fat reserves. [...] In 1972, a 3-month-old Gray Whale named Gigi was captured for brief study, and then released near San Diego.

	Relationship	Incorrect?
1	Grey whale – feed on – benthic crustaceans	

2	Grey whale – eat – benthic crustaceans	
3	Grey whale – turn on – grey whale's side	X
4	Grey whale – scoop up – sediments from the sea floor	
5	Grey whale – classified as – baleen whale	
6	Grey whale – has – baleen	
7	Grey whale – has – whalebone	
8	whalebone – act like – sieve	
9	baleen – capture – amphipods	
10	animal – feed in – northern waters	X
11	animal – live off – extensive fat reserves	X
12	ARG1 – capture – Gray Whale	X
13	ARG1 – release – Gray Whale	X

Table 1 - Relationships extracted from a Wikipedia article

Relationship 3) cannot be considered a general fact and sounds odd in isolation – using Carlson's (1977) terminology, the relationship refers to a stage of the whale and not to the individual. Relations 10) and 11) are incorrect as one would infer from them that they are applicable to all animals. Relations 12) and 13) are not incorrect as such but, considering that all the other relationships are about the kind 'Grey Whale', it would be tempting to believe that those similarly apply to the species – leading to false statements. We will not be discussing the problem of 3) in this work – and won't attempt to classify predicates. However, the 'genericity value' of the noun phrase is important to us: we want to know, for instance, when the information is about the kind 'grey whale' – as opposed to one particular whale – and to resolve referents as in the animal/whale example (referent resolution, as we will see later, is not limited to anaphora.) What we ideally want is a tool to annotate each noun phrase with its 'genericity value'.

In the next section, we will first give an overview of the issues linked to the automatic annotation of genericity, and show that the first step towards our end goal is to devise an appropriate, manual annotation scheme. The subsequent sections cover our attempts at designing such a scheme. We start by motivating our choice of labels for the annotation and go on to present our initial scheme, mostly based on few, intuitive questions. The issues encountered in using this scheme lead us to propose another, more complex scheme, which gives us better interannotator agreement. We finish with a short discussion of how our annotations correspond to particular ontological representations.

2. The automatic annotation of genericity: issues

A typical way to perform automatic linguistic annotations in computer science is machine learning. A program is given a corpus manually annotated by human experts and attempts to learn statistically significant rules which will then be tested on a separate corpus.

It is necessary to have a sufficiently large corpus, with a wide variety of examples, to perform such training. In the case of genericity, there is no corpus that we know of which would give us the required data. The ACE corpus (2005) is possibly an exception, but it only makes a distinction between generic and non-generic entities, which, as we will see later, is too vague for our purposes. The GNOME corpus (Poesio, 2000) is another example but it is limited in genres (the annotation guidelines are also specific to those genres) and again, it only has two genericity-related labels. It is therefore necessary to construct a separate training corpus for ontology extraction. Furthermore, manual annotation allows us to investigate distinctions that can be made on linguistic grounds and motivate them by individual examples, empirically

grounded by exhaustive annotation of corpus data. The use of multiple annotators leads to increased clarity and precision in such distinctions.

Annotating genericity, as we found out in our own initial attempts, is no trivial task. There are clear instances of non-generic entities such as proper nouns or narrative objects and clear instances as well of generic ones such as Latin species names, but when one starts considering every noun phrase in a corpus, things are far less obvious. Consider the following two sentences, taken again from a Wikipedia article:

(1) Later still, Hebrew scholars made use of simple monoalphabetic substitution ciphers.
(2) Cryptography has a long tradition in religious writing.

In the first sentence, it is not clear whether the article speaks of some Hebrew scholars or Hebrew scholars in general. Only detailed world knowledge of the topic would help us resolve this. In the second sentence, it may seem quite clear that the text talks about cryptography in general and it is probably difficult to imagine instances of the concept cryptography but there is only one entity in the world called cryptography and if one thinks of the difference between generics and non-generics as a matter of quantification, there is nothing that distinguishes the concept 'cryptography' from the Eiffel Tower. We would however want to argue that cryptography here is generic in the way that the sentence 'cryptographic messages have a long tradition in religious writing' refers to cryptographic messages in general. We found many other difficult examples in the course of the project.

Because it is difficult to rely on human intuition for this kind of annotation, we came to the decision that a precise scheme was needed, which would allow us to track the decision process made by humans when considering genericity, and which would eventually give us a quantifiable idea of how much agreement can be reached between two annotators.

3. Choosing labels

The first decision that we had to make was the choice of labels for the annotation. Krifka et al (1995) identify differences between genericity and non-genericity as well as between specificity and nonspecificity. Both object and kind-level entities can have a specificity value:

(3) A lion (as in 'A lion has a bushy tail') is non-specific and non-kind-referring. [1]
(4) Simba/a lion, namely Simba, is specific and non-kind-referring.
(5) A cat (in the taxonomic reading, as in 'a cat shows mutations when domesticated') is kind-referring but non-specific.
(6) The lion/a cat, namely the lion (taxonomic reading) is kind-referring and specific.

(Examples taken from Krifka et al, in Carlson and Pelletier, 1995).

As far as ontology creation is concerned, the difference between specific and non-specific readings for generic entities is not relevant. Theoretically, the information in (5) will be attached to the class 'cat1', that is, the node in the ontology which is parent to the nodes 'tiger', 'lion' and 'cat2', and whatever predicate comes after (6) will be linked to the class 'lion' in exactly the same way:

[1] Although we follow the Krifka et al general classification, we do not agree with the classification of this particular example. See our comments later in this section.

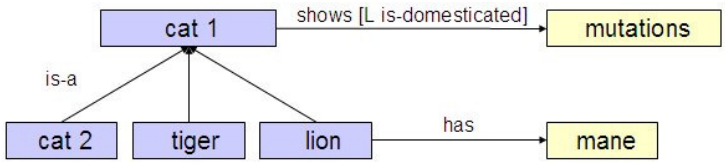

Figure 1 - A cat ontology

The notion of specificity is more interesting for object-level items, however, because if we attempt to merge nodes which refer to the same entity, only those nodes which are specific should be allowed to merge. See for instance the following three sentences (let's assume from the same text) and attached diagram:

(7) Flipper whistles on national TV.
(8) This dolphin has made a career!
(9) Mrs Smith has found a dolphin in her bath tub.

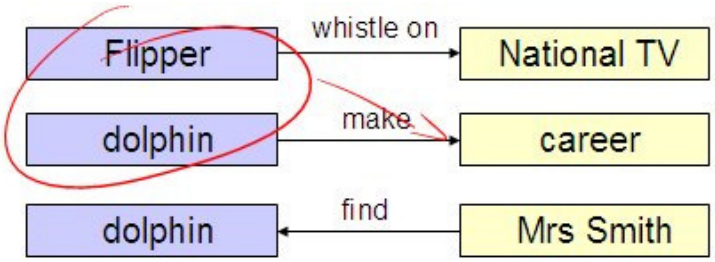

Figure 2 - A dolphin ontology

Here, only two out of three nodes should be allowed to merge, as the third dolphin is not specified and cannot be assumed to correspond to Flipper. From these initial distinctions, we get three possible labels:

– GEN: generic entities
– SPEC: non-generic, specific entities
– NON-SPEC: non-generic, non-specific entities.

We have mentioned in the course of the previous sections that some entities could be ambiguous between generic and non-generic (see the Hebrew scholars example) and that the referent of the noun phrase under consideration could be more specific than the lexical realisation appears to be (see the grey whale/animal example). We therefore introduced two further labels to deal with those cases:

– AMB: ambiguous between generic and non-generic
– GROUP: the text is referring generically to a subgroup of the noun phrase. So for instance, in the example of Section 1, the noun phrase 'the animal' refers to 'all of some' animals, namely all grey whales.

Finally, we also experimented with the idea that some concepts were not 'well established': the notion of a well-established concept comes from some work on definite singulars. Krifka et al (1995) show, for instance, that a definite singular can be used generically on a well-established kind only:

(10) The Coke bottle has a narrow neck.
*(11) *The green bottle has a narrow neck.*

(Examples attributed to Barbara Partee.)

We found the notion of well-established entity attractive from the point of view of ontology extraction, as we would like to be able to avoid the creation of dubious classes (for instance, we found in our training corpus the NP 'gimcrack affair', which in our view did not make for a stable concept with clear hyponyms).

It seems worth mentioning that when we are attributing a particular label to a noun phrase, we are only saying that the label applies to that noun phrase in context, i.e. to an instance of a grammatical construct such as bare plural or indefinite singular. In this respect, we are not making assumptions about the general features of any construct: in particular, when we are saying that a noun phrase is ambiguous between an existential and a generic reading, we do not mean to make a claim about whether bare plurals as a linguistic construct can be regarded as ambiguous or not. Similarly, if asked to annotate the noun phrase 'a lion' in 'a lion has a bushy tail', we would annotate it as a generic entity, even though Krifka et al (1995) argue that this is a non-kind-referring entity on the basis that the sentence is characterising and that the genericity does not occur at the level of the noun phrase. As far as this scheme is concerned, the lion in that sentence does not (necessarily) refer to a particular lion but rather to all lions, and we would therefore annotate it as generic.

4. Initial scheme

4.1. Instructions

Our initial scheme dealt with the distinction between generics and non-generics only, ignoring the problem of specificity. It used the four labels SPEC, GROUP, GEN and NON-WE. We give the complete scheme in Appendix 1 and comment next on the main points.

The non-generics were identified using a test of distinguishability: if an item X in the text can be distinguished from other Xs, then the noun phrase is probably at object-level (Step 1). The idea requires that unique individuals be dealt with separately in Step 2 (the Eiffel Tower cannot be distinguished from other Eiffel Towers, but it is nevertheless non-generic). We defined a unique object as a concept that doesn't have either children classes or instances. Well-established entities (Step 3) were identified with respect to their potential for being topics of information (could the noun phrase in the text be the headline of an encyclopaedia article?) Finally, groups were separated from real generics by considering the referent of the noun phrase (Step 4).

4.2. Results

The scheme was evaluated by presenting six annotators with four sections of the Wikipedia article 'The History of Cryptography', totalling 100 noun phrases, the boundaries of which were marked by one of the authors prior to annotation (this includes embedded NPs). The annotators were all students, with no formal knowledge of linguistics and with no previous experience of annotating genericity. Agreement was calculated for each pair of annotators according to the Kappa measure (Cohen, 1960):

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where Pr(a) is the actual, observed agreement and Pr(e) is the expected figure if annotators were acting independently and any agreement was due to chance. (There are different versions of Kappa depending on how multiple annotators are treated and how the probabilities

of classes are calculated to establish P(e): here we use the simple unweighted Kappa, Cohen 1960.) In practice, this means calculating the 'chance' confusion matrix from the annotation confusion matrix for a pair of annotators and then computing:

$$\kappa = \frac{Actual - Expected}{TotalAnnotations - Expected}$$

where *Actual* is the number of times that the two annotators agreed, *Expected* is the number of times that they would have agreed by chance, and *TotalAnnotations* is the total number of annotations performed. Note that taking chance into account means that there is no straight correspondence between plain percentages of agreement (*Actual /TotalAnnotations*) and Kappa values. Landis and Koch (1977) provide an interpretation of Kappa which is commonly used and that we summarise in Table 2.

Kappa	**Interpretation**
< 0	No agreement
0 - 0.2	Very low agreement
0.21 – 0.4	Low agreement
0.41 – 0.6	Moderate Agreement
0.61 – 0.8	Full Agreement
0.8 -1.0	Perfect Agreement

Table 2 – An interpretation of the values of Kappa

The evaluation gave us 15 Kappa figures ranging from 0.31 to 0.62, with percentage agreement ranging from 56% to 78%. Full results are available in Table 3. The figures in the top right-hand side triangle are the Kappas while the figures in the bottom left-hand side triangle are the plain percentages of agreement.

	1	**2**	**3**	**4**	**5**	**6**
1		.62	.40	.37	.51	.52
2	78		.39	.36	.48	.38
3	61	62		.31	.36	.45
4	60	61	56		.48	.43
5	68	67	57	65		.43
6	65	56	62	60	58	

Table 3 - Interannotater agreement , first scheme

4.3. Discussion

We carried out an analysis of the areas of stronger disagreement by picking out all NPs without majority. That is, we selected all phrases for which fewer than four of the six

annotators agreed on a label. We then attempted to understand the reason for each annotation and derived a rough map of issues.

We found that for 72 phrases out of 100, at least four annotators were in agreement. The remaining 28 phrases formed the core of our analysis. The following conclusions could be derived:

– A portion of the disagreements could have been covered by an 'ambiguity' label. In those cases, there is a genuine ambiguity between specificity and genericity, which is not resolved in the text. For instance, in the sentence *'Later still, Hebrew scholars made use of simple mono-alphabetic substitution ciphers'*, it is not clear whether all Hebrew scholars or some specific people are involved.

– Our simple definition of genericity in Step 4 ('all or any' of the instances of the concept expressed by the lexical item) produces confusions between GROUP and GEN labels. In the sentence *'The Greeks of classical times are said to have known of ciphers'*, 'ciphers' was labelled by some annotators as GROUP, presumably because the Greeks would not have known of all ciphers in existence.

– The semantic interpretation of the copula differs depending on the annotator. Some assume that the second argument of the copula has automatically the same scope as the first argument, while others prefer to see it as a class (and therefore a generic entity). An example of such disagreement occurs for the NP 'examples of cryptography' in the sentence *'Herodotus tells us of secret messages [...] these are not proper examples of cryptography.'* Out of the two annotators who marked 'secret messages' as specific, one marked the last NP as generic, the other as specific.

– In some cases, the disagreement is simply due to slight differences in the resolution of the referent in context, as in *'Gilbert Vernam proposed a teletype cipher in which a previously-prepared key [...] is combined character by character with the plaintext message.'* When annotating the NP 'the plaintext message', opinions differed between SPEC and GEN labels (with an odd GROUP), presumably because of interpreting the sentence as either an exemplary, specific event or as generic instructions.

The last point deserves some elaboration. It is interesting to note that all the issues above relate, in one way or another, to the resolution of the referent in context. Either the annotators do not realise the ambiguity of the noun phrase, or they disagree on the resolution. This accounts for the confusions between group and generic labels: in the previous example, *'The Greeks of classical times are said to have known of ciphers'*, some people resolved 'ciphers' to 'all ciphers at all times' while others interpreted it as 'ciphers in ancient Greece'. The issue relating to the copula can be similarly interpreted, where the second argument is seen as referring to the first one in one case, or to a class of objects in the other case. In order to track this problem, we attempted in further schemes to give rules on how the referent should be resolved and asked the annotators to provide a written record of each noun phrase's referent. The issue is further discussed in Section 5.

Finally, we found that the notion of 'well-established' entity was particularly difficult to explain and to exemplify in a structured scheme – it proved to be the source of many preliminary queries from our annotators. Note that there are also now arguments against the idea of well-established kinds on the grounds that sentences starting with a definite article such as the following are possible (Hofmeister, 2003):

(12) The newly-hatched fly is a lazy insect.
(13) The well-crafted bottle has a narrow neck.

Considering the mediocre results given by this label, and the lack of strong basis for its linguistic motivation, we abandoned it in subsequent experiments.

5. A revised scheme

5.1. Instructions

We found that our Wikipedia article, although giving us a fair proportion of both specific and generic instances, was lacking many types of expression that we had noted elsewhere in text. We suspected that this was due to the set encyclopedic style of Wikipedia and turned to a different corpus. The new scheme was developed using material which would give us better examples of the range of expressions found in general text: we produced a development corpus by selecting 10 paragraphs out of 10 different genres in the British National Corpus and marking 50 noun phrases for their variety of expressions. This corpus is available at http://www.cl.cam.ac.uk/~ah433.

Having gone through several development iterations, our scheme now contains 14 steps and caters for specific cases such as existentials, proper nouns and copula constructions. We give the full scheme in Appendix 2 and discuss here the main points.

– We have now a step dedicated to referent resolution (Step 4). The annotator is required to perform not only simple anaphora resolution but also spatial and temporal resolution in context. Pronouns, and in particular possessive pronouns, must refer to an entity in the text.

– Unique entities are dealt with in two steps (7 and 8), one to assert the uniqueness of the noun phrase and the other to filter through class names which could be interpreted as unique objects: our initial experiments, for instance, showed that the concept of 'cryptography' had been marked as specific by some annotators because 'there are not several cryptographies' (see Section 3).

– A label for non-specifics has also been added (Step 12): our initial definition of generics as 'any' or 'all' of a class instances created problems when annotating sentences such as 'I want a new bike', where 'bike' is 'any bike' but certainly not a generic entity. We now require that entities that refer to a particular object be classified as either specific (identifiable) or non-specific (non-identifiable). We explain this distinction further at the end of the section.

– The differentiation between groups and generics is now made simple by comparing the textual entity P with its referent resolution P2: when P = P2, we have a generic entity, otherwise a group (Step 14).

– Finally, there is an explicit ambiguity label which can be applied to bare plurals (Step 15). Non-specific entities in bare plurals must be reconsidered and the annotator is asked whether there is a reading of the sentence where the entity might refer to a class of objects.

Some comments must be made about the tests for specificity and non-specificity. As argued by Jorgensen (2000), specificity is not a well-defined concept. The idea behind the notion is that specific entities are identifiable while non-specific ones are not. Jorgensen, however, quotes Krifka et al (1995) to show that there is no good consensus on what the definition actually is:

> The actual specific/non-specific distinction (if there is just one such distinction) is extremely difficult to elucidate in its details. It is for this reason that we wish to remain on a pretheoretic level. Even so, we had better point out that we take, e.g., a lion in 'A lion must be standing in the bush over there' to be specific rather than nonspecific, even if there is no particular lion that the speaker believes to be in the bush.

Jorgensen himself proposes a definition centred on the speaker: what he calls J-specificity separates the cases where the speaker has the means to identify the referent and/or believes it to be unique from cases where neither necessarily apply. The latter cases are non-specific. We adopted that approach and chose to have a test for specificity after the test for distinguishability: if an entity can be distinguished from other similar entities (ie, if it is unique in Jorgensen's sense) and if it is identifiable, then it is specific; if the entity can be distinguished from other similar entities but is not identifiable, then it is non-specific.

5.2. Results

The test corpus comes from the BNC, like the development corpus (see Section 5.1). Our software randomly selects 10 different genres in the BNC, randomly extracts one paragraph for each genre, produces a syntactic parse of each sentence using the RASP software (Briscoe and Carroll, 2000) and identifies full NPs in the output. At this stage, it is necessary to manually weed incorrect NPs due to parsing errors. In our first trial, we found that out of 552 noun phrases, 131 were incorrect, yielding an accuracy of 76%. We also decided that when two NPs were co-ordinated, they should be considered in isolation.

In order to measure Kappa on the new scheme, we extracted the first 48 noun phrases out of the 421 left after parsing and manual correction, and presented them to two annotators. We obtained a Kappa of 0.744, corresponding to an agreement of 83%. The confusion matrix for this annotation is shown in Table 4.

	SPEC	**NON-SPEC**	**GEN**	**GROUP**	**AMB**
SPEC	25	0	0	0	0
NON-SPEC	0	6	0	2	0
GEN	0	0	3	1	1
GROUP	0	2	2	6	0
AMB	0	0	0	0	0

Table 4 - Confusion matrix for final annotation

5.3. Discussion

Most of the disagreements left at this stage relate to the resolution of the referent. See for instance the sentence:

(14) Under the agreement, enhancements to the libraries will be developed to address such areas as performance, ease-of-use, internationalisation and support for multi-threading.

One annotator resolved 'enhancements to the libraries' as 'enhancements to the Tools.h++ libraries[2]' while the other resolved it to 'enhancements that will be developed under the agreement', producing a NON-SPEC annotation in the first case and a GROUP annotation in

[2] For non-computer scientists: the Tools.h++ library is a piece of software designed for programmers who write in the C++ language, mentioned in the text prior to sentence (14).

the second case. We found that disagreements were often due to differences in world knowledge, simple omissions, and interpretation of the semantics of certain verbs. For instance, in a sentence of the type 'X consists of Y', some people tend to resolve Y as being 'the Y in X' while others will resolve it as just Y. This corresponds roughly to the two paraphrases 'X consists of a certain amount of Y' and 'X is made out of the general material called Y'.

We expect reference resolution to be one of the major problems in the design of a program for the annotation of genericity, as automatic reference resolution beyond anaphora is currently limited (see Vieira and Poesio, 2000 for a description of the difficult problem of processing bridging descriptions). We argue, however, that considering the referent is absolutely necessary to provide accuracy to the results. As a short investigation of how the referent affects the annotation, we decided to perform an experiment where the same 48 noun phrases would be annotated, first with and then without the reference resolution step. The results showed major differences: the group labels, which we would have ideally liked to see transferred to generic labels, were transferred to specifics and non-specifics. Out of 48 noun phrases, it was judged that the annotation for 15 of them was incorrect beyond argument when ignoring the reference resolution step.

6. Interpreting genericity

We have so far related the phenomenon of genericity to a basic ontological structure, with classes and instances. For instance, a generic will refer to a concept, or class, which can have instances. An entity marked as specific will be taken as being an instance of a class in the ontology. Some specifics will even be marked as unique instances of that concept: for example, there is only one instance of the concept 'Great Wall of China'. Non-specifics will also refer to instances, but those will be random instances of a concept rather than identifiable ones. In the end, the various annotations can be paraphrased as follows:

– X is specific: X is an instance of a concept (or unique instance of that concept, depending on the path followed for the annotation)
– X is non-specific: X is a random instance of a concept
– X is generic: X is a class
– X is a group: X refers to some instances of a class which themselves form a subclass.

The problem that has not yet been explored in this work is that of providing a particular interpretation of genericity such as, for instance, quantification, rules or induction (see Cohen, 2002). The reason that such interpretations cannot be directly linked to the notions of quantification or inference that they appeal to is that generic statements differ in the extent to which they apply to individuals in a class. See for instance:

(15) Turtles are reptiles. (All turtles are reptiles.)
(16) Turtles lay eggs. (Female turtles lay eggs.)
(17) Turtles live over 100 years. (Some turtles, in rare cases, live over 100 years.)

It is clear that examples (16) and (17) cannot be treated as involving universal quantification over individual turtles. However this is a highly complex issue with no clear solution which we will not discuss further here.

6. Conclusion

This paper has presented a scheme for the manual annotation of genericity, with the ultimate aim of using human annotations to train an automatic classifier. Our current scheme produces reasonable interannotator agreement with a Kappa of 0.74. We noted that the resolution of

the noun phrase's referent has much to do with the way it is annotated by humans and we predict that this might be the main bottleneck in automating the annotation. Finally, we showed how the annotations could relate to the concept nodes of an ontology and remarked that the notion of generic class would have to be further formalised to take into account all possible interpretations of genericity. We leave this problem as further work, together with the construction of a machine learner able to automatically reproduce human annotations.

References

ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 5.6.1 2005.05.23. 2005.

Briscoe, Ted; Carroll, John and Watson, Rebecca.
2006 The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia, 2006.

Carlson, Greg.
1977 A Unified Analysis of the Bare English Plural. In: *Linguistics and Philosophy.* **1**:413-457.

Carlson, Greg and Pelletier, Francis.
1995 *The Generic Book.* University of Chicago Press.

Cohen, Axel.
2002 Genericity. In *Linguistische Berichte*, 10.

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

Hofmeister, Philip.
2003 Generic Singular Definites. Available at http://www.stanford.edu/~philiph/skeleton.pdf. Last accessed on 24 April 2008.

Jorgensen, Stig.
2000 Computational Reference. Ph.D Dissertation, Copenhagen Business School.

Krifka, Manfred et al.
1995 Genericity: An Introduction. In G. Carlson and F. Pelletier, eds., *The Generic Book*. Chicago: University of Chicago Press.

Landis, J. R. and Koch, G. G.
1977 The measurement of observer agreement for categorical data. In: *Biometrics*. Vol. 33, pp. 159—174.

Poesio, Massimo.
2000 Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proceedings of the Third Conference on Language Resources and Engineering*, Athens.

Vieira, Renata and Poesio, Massimo.
2000 Corpus-based development and evaluation of a system for processing definite descriptions. In *Proceedings of the 18th COLING*, Saarbruecken, Germany.

Appendix 1

The annotation scheme is designed as a list of 4 steps, to be taken in turn. Annotators should start at Step 1 and follow instructions in each subsequent step. When one of the four labels SPEC, GEN, GROUP or NON-WE appears, the noun phrase should be marked appropriately and the annotator should stop.

Step 1. Can the entity be singled out from similar entities in the real world? If yes, annotate as SPEC.
Examples:
[The development of cryptography] has been paralleled by the development of cryptanalysis: 'the development of cryptography' cannot be separated from other potential 'developments of cryptography' in this context, so go to next step without annotation.
[Many scholars] believe it's a concealed reference to the Roman Empire: those 'many scholars' can probably be distinguished from other scholars, so annotate as SPEC.
[The Chinese] invented the firework: There is only one Chinese people and it doesn't make sense to talk about another entity called 'the Chinese', so go to next step without annotation.

Step 2. Is it possible to imagine instantiations or specialisations of the expression (is the expression a class)? Try to instantiate the phrase P with 'this P' (or 'these' if plural), or 'this form of P', which produces another referent. If the instantiation does not make sense, annotate as SPEC.
Examples:
Until the 1970s, [secure cryptography] was largely the preserve of governments: 'this form of secure cryptography' is acceptable, so go to next step without annotation.
The breaking of [codes] and ciphers: 'this code' is acceptable, so go to next step without annotation.
Allied reading of Nazi Germany's ciphers shortened [World War II]: 'this World War II' or 'this form of World War II' is nonsensical, so annotate as SPEC.

Step 3. Is the expression a 'well-established entity'? I.e. can you define the entity in a way that most people would agree with, or could you imagine an encyclopaedia article about it – or at least a webpage on the topic? If not, annotate as NON-WE
Examples:
The subsequent introduction of electronics and computing has allowed [elaborate schemes of still greater complexity]: it is difficult to imagine what an article about 'elaborate schemes of still greater complexity' would look like, so annotate as NON-WE.
[Methods of encryption that use pen and paper]: although very constrained, this is a definable topic, so go to next step without annotation.
Methods of encryption that use [pen] and paper: 'pen' is a concept that would have a definition in a dictionary, so go to next step without annotation.

Step 4. Can the information contained in the text apply to the entity in general (that is, to all or any of its instances), or is it only relevant to a subgroup? If yes, annotate as GEN, otherwise, annotate as GROUP.
Examples:
It's a concealed reference to the Roman Empire (and so to [Roman persecution policies]): this is a reference to all Roman persecution policies, therefore annotate as GEN.
[The key] in every such system had to be exchanged between the communicating parties: this is not applicable to all keys, or any key, but to certain keys in certain systems, therefore annotate as GROUP.
The destination of the whales is California where they breed and [the young] are born: this is not applicable to all young but only to all whale youngs, therefore annotate as GROUP.

Appendix 2

The annotation scheme is designed as a list of 14 steps. Each step corresponds to a test, the answer to which decides of the next step to take. Annotators should start at Step 1 and follow instructions in each subsequent step. When one of the five labels SPEC, NON-SPEC, GEN, GROUP or AMB appears, the noun phrase should be marked appropriately and the annotator should stop unless a further step is specified.

In what follows, the letter P refers to the noun phrase being annotated.

1. Does P appear in an existential construct of the type 'there [be] (a) P(s)' where the existential describes a particular situation in space and time? The copula 'be' can appear conjugated in any tense and there may be other phrases separating it from its logical subject and object. Yes –> SPEC No –> 2.

Examples: *There are daffodils in the garden. There is a car parked across the road.*

2. Is P a proper noun? Yes –> SPEC No –> 3.

Examples: *John Smith, War and Peace, Easter Island, World War II...*

3. Does P start with an indefinite quantifier other than 'all' or 'a'? This includes some, a few, few, many, most, one of, a couple of, etc. Yes –> NON-SPEC No, quantifier is 'all' –> read 4 and jump to 13 No –> 4.

4. It will often be the case that P refers to something more specific than it seems. We will call that more specific reading P2. Consider the following cases:

Make sure you have a hammer at hand. The tool will be needed in step 4. The tool = the hammer.

The cathedral is magnificent. The stained-glass windows date from the 13th century. The stained-glass windows = the stained-glass windows of the cathedral X.

(In a book about 19th century Europe) *Women could not vote.* Women = women in 19th century Europe.

It helps to ask whether the entity could be 'any P in existence'. If not, there is probably a more specific reading involved (P2). Sometimes, it is difficult to answer the question (for instance if the entity is not a familiar concept). In this case, try to specify as much as possible using explicit location/time details from the context. If the word is an anaphoric reference to a previous word, let P2 be the previous word. If the identity of P can be specified, through context or general knowledge, record the specified entity in P2. Possessives should also be resolved:

his car = Paul's car.

If P cannot be precised further, then P2 = P. Go to step 5.

5. Does P appear in a construct of the type 'A [be] P(s)'? The copula be can appear conjugated in any tense and there may be other phrases separating it from its logical subject and object. Implicit copulas also call for an affirmative answer: for instance, 'X classified as Y' or 'X named Y' will, in certain cases, mean 'X is a Y.' Yes –> 6 No –> 7.

6. In the identified construct, 'A [be] P (s)', are A and P two names for the same thing? Consider for example: Elizabeth II is the Queen of England or The Morning Star is the Evening Star. Yes –> 7 No –> 13.

7. Does the lexical realisation of P2 refer to an entity which is unique in the world? (Plurals are necessarily non-unique.)Yes –> 8 No –> 9. (If unsure, go to 9.)

The Daily Mail was looking for a new chief editor. Paul went for the job. There are not several jobs of chief editor for the Daily Mail, so this is unique.

The lion bit my toe. There is more than one lion in the world, possibly even more than one lion who bit my toe, so this is not unique.

8. Is P2 a common noun that could have taxonomical children? (When considering complex entities, e.g. 'X of Y', the taxonomical child must belong to the head of the phrase – X in the example.)Yes –> 13 No –> SPEC.

Psychology: yes, because experimental psychology and behavioural psychology are forms of psychology.

Mozart's death: no, nothing is a form of Mozart's death.

9. Does P have a determiner? Yes –> 10 No –> 11.

10. Does P2 refer to a particular object, or group of objects, in the world? Ie, out of all possible P2s, is the text only talking about one/some of them? (The determiner is sometimes a very good clue.) Yes –> 12 No –> 13.

Can you see the lion? (Assuming P2=the lion at London zoo.) Particular object = the lion being pointed at, as opposed to all possible lions at London zoo.

The lion is a mammal. Non-particular = this is talking about lions in general.

11. Is there a reading of the sentence where the P2(s) in the text can be distinguished from other P2s (or group of P2s)? Ie, out of all possible P2s, is the text only talking about some of them? Yes –> 12 No –> 13. (See examples in step 10.)

12. Is/Are the P2(s) in the text identifiable or is the text talking about, potentially, any of them? Identifiable –> SPEC Not identifiable and last step = 10 –> NON-SPEC Not identifiable and last step = 11 –> NON-SPEC + 14.

Mary has a new bike. Identifiable = one bike, Mary's bike.

I would like a new bike. Not identifiable = one bike, but any will do.

13. Is P2 = P? Yes –> GEN No –> GROUP.

14. Is there a reading of the sentence where P2 means 'all' P2(s)? Yes –> AMB No –> STOP.