Technical Report

UCAM-CL-TR-795
ISSN 1476-2986

Number 795

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Underspecified quantification

Aurelie Herbelot

February 2011

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted 2010 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Trinity Hall.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Underspecified quantification

Aurelie Herbelot

Summary

Many noun phrases in text are ambiguously quantified: syntax doesn't explicitly tell us whether they refer to a single entity or to several and, in main clauses, what portion of the set denoted by the subject Nbar actually takes part in the event expressed by the verb. For instance, when we utter the sentence *Cats are mammals*, it is only world knowledge that allows our hearer to infer that we mean *All cats are mammals*, and not *Some cats are mammals*. This ambiguity effect is interesting at several levels. Theoretically, it raises cognitive and linguistic questions. To what extent does syntax help humans resolve the ambiguity? What problem-solving skills come into play when syntax is insufficient for full resolution? How does ambiguous quantification relate to the phenomenon of genericity, as described by the linguistic literature? From an engineering point of view, the resolution of quantificational ambiguity is essential to the accuracy of some Natural Language Processing tasks.

We argue that the quantification ambiguity phenomenon can be described in terms of underspecification and propose a formalisation for what we call **underquantified** subject noun phrases. Our formalisation is motivated by inference requirements and covers all cases of genericity.

Our approach is then empirically validated by human annotation experiments. We propose an annotation scheme that follows our theoretical claims with regard to underquantification. Our annotation results strengthen our claim that all noun phrases can be analysed in terms of quantification. The produced corpus allows us to derive a gold standard for quantification resolution experiments and is, as far as we are aware, the first attempt to analyse the distribution of null quantifiers in English.

We then create a baseline system for automatic quantification resolution, using syntax to provide discriminating features for our classification. We show that results are rather poor for certain classes and argue that some level of pragmatics is needed, in combination with syntax, to perform accurate resolution. We explore the use of memory-based learning as a way to approximate the problem-solving skills available to humans at the level of pragmatic understanding.

Acknowledgments

I would like to thank....

... my supervisor Dr Ann Copestake, who supported me in more ways than will fit on this page: in particular for mentioning the phenomenon of genericity in the first place, for reading and re-reading countless papers and drafts and always providing the most enlightening comments on my work, for pushing me in the right research direction at times of confusion and for leading me to believe that computational linguistics is, all in all, a rather enjoyable occupation.

... my partner Eva, who had to patiently listen to countless bad ideas of mine and often found the flaws in my argument. For three years of commuting between Berlin and Cambridge, and travelling with me to faraway places. For repeating in persuasive terms that this thesis would, despite contrary beliefs of mine, see the light of day.

... Dr Simone Teufel for invaluable guidance on the topic of annotation, for guiding me through the design of both annotation guidelines and experiments and for precious comments on the relevant chapter in this thesis.

... Mohan, for many theoretical discussions, for reading drafts in the most thorough fashion I know and for teaching me the maths I never knew. Also for hours of semi-idleness in the cosiest room of all Cambridge. For being there for me (us) in thorny situations.

... Diarmuid, for highly educational conversations about annotation and machine learning. For answering many e-mails about SVMs.

... all those who took part in tiresome annotation efforts: Mohan again, the friends of the Sweden house (particularly Stefan, who made it through to the final experiment) and Johanna in the NLP group.

... my friend Joanne for her support and wisdom at times of critical deliberation. I would not have achieved what I have without her.

... Matthias and Christian, who think that computational linguistics is ‘cool’, always ask so much and make me think properly.

... Claudia and Nurettin, for providing me with a quiet working space in the last months of my PhD.

... Iwona, who always thinks that everything will be fine and whom I sometimes believe.

Finally, I would like to thank both my examiners, Dr Stephen Clark and Dr Carl Vogel, for their thorough comments on this work. I hope I have done justice to them in revising the text of this thesis.

Contents

1 Introduction **11**

1.1 Quantification for NLP 12

1.1.1 Extracting facts 12

1.1.2 Performing inference at instance-level 13

1.1.3 Inference via entailment 14

1.2 Previous and related work 15

1.3 Objectives 19

1.4 No objects 20

1.5 Outline 22

2 A theoretical account of ambiguously quantified noun phrases **23**

2.1 Terminology 24

2.1.1 Quantification resolution 24

2.1.2 Quantifiers 26

2.1.3 Scoping the study 28

2.1.4 Reference: sets and lattices 30

2.2 From bare plurals to ambiguous quantification: the genericity phenomenon 31

2.2.1 Bare plurals 31

2.2.2 Genericity: some definitions 32

2.2.3 The GEN operator 36

2.2.4 Semantics of the generic sentence 39

2.2.5 Interaction with other linguistic phenomena 42

2.2.6 Genericity in psychology 44

2.3 Definite plurals: in or out? 45

3 Underspecified quantification **47**

3.1 Link's notation (1983) 47

3.2 Bare forms, true kinds and wide-scope stereotypes 49

3.2.1 Bare plurals and underspecified quantification 49

3.2.2 True and derived kinds 54

3.2.3 Wide-scope stereotypes 55

3.2.4 Are bare singulars bare (or even singular)? 57

3.3 Definite plurals included 58

3.4 Formalisation 61

3.4.1 Formalising collective and distributive predicates 61

3.4.2 Formalising kinds and stereotypes 64

3.4.3 Formalisation summary 67

3.4.4 A remark on semantics 68

4 Quantification resolution, the human way **69**

4.1 Linguistic annotation: motivation and theory 70

4.1.1 Linguistic motivation 70

4.1.2 NLP motivation 72

4.2 Annotation corpus 73

4.2.1 Parsing pipeline 75

4.3 Evaluation of annotation agreements 78

4.4 An annotation scheme for quantification resolution 79

4.4.1 Theoretical claims: a reminder 79

4.4.2 Scheme structure 80

4.4.3 Material 81

4.4.4 Definitions 81

4.4.5 Guidelines and decision trees 83

4.5 Implementation and results 85

4.5.1 Task implementation 85

4.5.2 Kappa evaluation 86

4.5.3 Annotation issues 91

4.6 Advantages over genericity annotation 94

5 Automating quantification resolution **97**

5.1 The gold standard . . . 97

5.1.1 Building the gold standard . . . 98

5.1.2 Class distribution . . . 99

5.2 A syntax-based classifier . . . 99

5.2.1 Some theory . . . 99

5.2.2 Features . . . 101

5.2.3 The classifier . . . 102

5.2.4 Experimental setup . . . 103

5.3 Results and discussion . . . 103

5.3.1 Results . . . 103

5.3.2 Some correlations with linguistic theory . . . 105

6 Quantifying with similarity **109**

6.1 Where is the quantification constraint? . . . 109

6.2 Situational analogy . . . 111

6.2.1 Which similarity? . . . 112

6.3 Situational analogy: system implementation . . . 115

6.3.1 The data . . . 115

6.3.2 The similarity measure . . . 115

6.3.3 The situational analogy algorithm . . . 119

6.4 Results . . . 119

6.5 Issues with nearest neighbour algorithm . . . 126

6.6 Beyond nearest neighbours? . . . 128

7 Conclusion **133**

7.1 Contributions . . . 133

7.2 Quantification resolution in the real world . . . 134

7.3 Remaining issues . . . 135

A Guidelines for the quantification annotation task **147**

A.1 Material 147

A.2 The task 147

A.2.1 Some definitions 148

A.3 More on quantification 148

A.3.1 The annotation labels 149

A.3.2 What to do when you hesitate? 150

A.4 The annotation process 150

A.4.1 Quantified NPs 150

A.4.2 Proper nouns 151

A.4.3 (Non-bare) singulars 151

A.4.4 Plurals 152

A.4.5 Bare singulars 152

A.5 Decision trees 153

B Subset of annotated data **155**

Chapter 1

Introduction

This thesis is ultimately about reference — or rather about referents, that is, about the *things* that we talk about when we use noun phrases such as *the cat*, *God* and *mosquito bites*. More exactly, it concerns itself with the quantities implied by such noun phrases, i.e. not with cats and God directly but with numbers of cats and numbers of God(s).

In the course of this work, we will attempt to elucidate, for instance, how many cats are referred to in the sentence *The cat is sleeping by the fire*. The answer to this question may seem obvious but this assumed clarity is only due, as we will show, to the sophistication of our reader's language skills. The noun phrase *the cat* is actually highly ambiguous in essence. Let us imagine a biologist writing, in an encyclopaedia article, *The cat is a mammal*. The topic of the sentence denotes many more entities than if the biologist mentions the same phrase, *the cat*, in her living room, uttering the above sentence: *The cat is sleeping by the fire*. The former allows us to deduce that given a random cat, this cat is a mammal, while the latter certainly does not imply that all cats always sleep by the fire in the biologist's living room — as much as they would like to. The quantity expressed by the noun phrase is in some sense hidden.

The object of this work is to retrieve the quantities alluded to by ambiguous noun phrases. We will call this process **quantification resolution**, or in short, **quantifying**.

There will be very little about the reference phenomenon itself in the following pages: in particular, we will happily avoid the numerous philosophical and linguistic debates brought about by the concept. Whether noun phrases refer to actual things in the world or to ideas, what this means for dead people, and whether Sherlock Holmes must be made to exist to utter the sentence *Sherlock Holmes does not exist* is of no concern to us. We take the stance that noun phrases do refer and that, regardless of what they refer to, they can be seen as quantifying: the word *cat* in a particular sentence may denote real flesh and blood cats or merely ideas of cats, but those cats or those ideas will be one or several or many or all. Further, we can often resolve the initial ambiguity by generating signs (or more specifically signifiers in the Saussurian sense) that are both explicit in terms of

quantification and logically compatible with all possible agreed referents for our sentence — this, without ever considering what those referents might be:

1. **All** cats are mammals.
2. **One** cat is sleeping by the fire.

The subjects in Sentences 1 and 2 are now differentiated. We will say that they have been **explicitly quantified**. Note that we haven't paused to specify the meaning of *cats*. We simply assume, in the tradition of model-theoretic semantics, that noun phrases have an extension in the actual world and that this extension is shared among speakers. For example, when people utter the sentence *Sharks are dangerous* (meaning *Most/All sharks are dangerous*), they only take into account currently living sharks and ignore the dead and unborn animals which, presumably, would change the meaning of the sentence into *Some sharks are dangerous*.

In the rest of this work, we posit (unless otherwise indicated) that quantification resolution can be regarded as paraphrasing and studied at the level of the lexical sign. The resolution process has the side-effect of providing quantification for the referent itself, but we will not discuss the computational means to obtain a representation of that referent.

1.1 Quantification for NLP

In what follows, we will show that quantification resolution is essential for the performance of various Natural Language Processing tasks. We will first focus on the automatic construction of factual databases and then consider inference operations that can be effected over such databases. We will argue that, both in the process of building such resources and in the process of using them for AI-related tasks, recording quantification values is necessary to the accuracy of the world model that they offer.

1.1.1 Extracting facts

Consider the following paragraph, taken from an article in the online encyclopaedia Wikipedia[1]:

> "The Four-toed Hedgehog (Atelerix albiventris), or African Pygmy Hedgehog, is a small species of hedgehog found throughout much of the south-Saharan African countries, from Senegal and Mauritania in the west, to Sudan in the east, and it has been recorded as far south as Zambia. [...] The Four-toed Hedgehog [...] has

[1] http://www.wikipedia.org/, last accessed 16th August 2010.

> short legs, a long nose, and small beady eyes. It can vary greatly in coloration [...] When the Four-toed Hedgehog is introduced to a new or particularly strong smell, it will sometimes do what is referred to as self-anointing. It creates a frothy saliva and spreads it onto its quills in incredible amounts. It is not really understood why it does this, but it is thought to be a defensive action, as hedgehogs have been known to self-anoint with poisonous toads. [...] The Four-toed Hedgehog [...] is even displayed in competitive hedgehog shows."
>
> (Wikipedia. 'The Four-toed Hedgehog'. Accessed 26th January 2010.)

For the sake of the example, we will imagine that some information extraction software has retrieved the Wikipedia page as part of an effort to construct a biological database. The software is supposed to pick out general facts about different animal species and integrate them in an electronic resource. Without much knowledge of quantification, the system might assume that all definite singular noun phrases refer to single individuals, as in *The cat is sleeping by the fire*. The resulting analysis of the Wikipedia article would then inform the reader of the eventually created resource that the particular hedgehog under consideration — let's call him Harry — is an outstanding runner. It has been observed from Senegal and Mauritania to Sudan, and as far as Zambia. It will come as no surprise that Harry is displayed in competitive hedgehogs shows.

Similarly, the system might decide that bare plurals refer to whole classes, as in *Cats are mammals*. Renewed reading of our database would now indicate that all hedgehogs sometimes cover themselves with poisonous toads.

Those decisions are actually well motivated. It is statistically true that definite singular noun phrases overwhelmingly refer to individual, specific entities while bare plurals are slightly more likely to refer to a majority reading (paraphrasable via *most* or *all*) than to an existential *some* reading (we list some grammatical constructions and the distribution of their possible readings in Section 5.2.1). Unfortunately, as will be shown throughout this thesis, the statistics quickly break down — and lead us into scenarios where hedgehogs commonly transport dead toads on their backs. We thus argue that quantification resolution is necessary in information extraction.

1.1.2 Performing inference at instance-level

We have just shown that performing adequate quantification resolution when processing text for information extraction would result in improved accuracy in the produced resources. We will now demonstrate that it would also lead to increased precision in the inferences computed from such resources.

We will assume a task where it is desirable to make inferences about instances of a certain concept. We will also make the simplifying assumption that there is a direct mapping between the quantification of the statement involving that concept and the likelihood of

its instances to engage in the situation described by the statement. Then, if a group of instances is distributionally quantified via q (a quantifier such as *some* or *most*) in relation to a predicate vp, then each instance in that group has a probability q to take part in vp. We will ignore here the problem of defining p for each q and express such inference using probability adverbs. See for instance:

> IF *all* four-toed hedgehogs are hedgehogs AND Harry is a four-toed hedgehog $THEN$ Harry is *definitely* a hedgehog
>
> IF *most* four-toed hedgehogs have short legs AND Harry is a four-toed hedgehog $THEN$ Harry *probably* has short legs
>
> IF *some* four-toed hedgehogs are displayed in competitive shows AND Harry is a four-toed hedgehog $THEN$ Harry is *possibly* displayed in competitive shows

In order for this mechanism to function, note that the premises must be appropriately quantified. Without this, inference is impossible.

1.1.3 Inference via entailment

We will now turn to another issue related to inference. Let us imagine that we have a database, or ontology, which contains the relation *Mary – has – Siamese*. One desirable feature of the query mechanism would be that, given the user question *Does Mary have a cat?*, the system would reply affirmatively, having made the inference that *Siamese – is a – cat*, or in other terms that *Siamese* entails *cat*.[2]

3. (a) SYSTEM: Mary – has – Siamese
   (b) USER: Does Mary have a cat?
   (c) SYSTEM: Siamese – is a – cat Mary – has – cat

The assumption usually made about entailment at the word level is that it is mostly a lexical problem. The task of finding pairs of lexically entailing words has recently received much attention in the literature. It is usually subdivided into two subtasks:

> finding potential replacements for a given word (usually by considering the distribution of that word in large corpora: see Lin, 1998; Szpektor et al, 2004)

[2]Most computational linguists (e.g. Giuliano and Gliozzo, 2007) assume a link between taxonomy and entailment which is apparent in many cases of entailment between nouns. Croft and Cruse (2004) give a definition of hyponymy based on entailment — but also show the limits of such an approach: *Basil became a Catholic* does not entail *Basil became a Christian* and *The wasp stung John on the knee* entails *The wasp stung John on the leg*, despite the fact that *Catholic* is a hyponym of *Christian* and *knee* is no hyponym of *leg*.

deciding whether a candidate replacement fits the original word in a particular context (see Dagan et al, 2006).

The literature tells us that if *Siamese* entails *cat*, we can substitute the latter for the former as long as the senses of the two words in context match:

4. Mary's Siamese Mary's cat.

However, consider the following:

5. All Siamese (have blue eyes) $\not\Rightarrow$ All cats (have blue eyes).

The substitution is this time not possible, not because of a sense mismatch but because of the quantification of the noun. The extension of *Mary's Siamese* and *Mary's cat* is the same but *all Siamese* and *all cats* refer to two different sets of individuals, making the entailment impossible. That is, entailment via word substitution only works over restricted, usually existentially quantified, sets. This demonstrates the need for quantifying constructs that are by essence ambiguous:

6. Siamese have blue eyes $\not\Rightarrow$ Cats have blue eyes

Note that the sentence *Cats have blue eyes* is actually true in a context where, for instance, two people are arguing whether cats can ever have blue eyes. But its correct paraphrase is then *Some cats have blue eyes*. The entailment is safe at the surface level (it is true that if all Siamese have blue eyes, then some cats have blue eyes) but not at the formal level where the premise must be universally quantified and the conclusion, after word substitution, preserves the universal quantifier.

1.2 Previous and related work

Section 1.1 informally illustrated how quantification resolution is necessary to the accuracy of some NLP tasks. In this section, we report in a more formal way some related work and situate our task in the field.

Our work on quantification stems from research in the field of ontology extraction. In Herbelot and Copestake (2006), we showed that it was possible to extract (biological) taxonomic relationships from Web data with very high precision, using semantic parsing. The reason we obtained good results in that task is that the problem was well-contained: we were using a small number of clear patterns to extract general is-a relationships, which were then filtered according to the lexical nature of the subject and object in the sentence (we only kept those relationships that involved species names). Trying a similar method

on a slightly harder task — returning the typical food of various animal species — proved a lot less successful. Despite the corpus being restricted to encyclopaedic data (the online resource Wikipedia), we nevertheless returned relations such as 'cats eat chocolate' and 'dogs eat pudding'. The reason for this was the assumption that all plurals are universally quantified[3].

The consequence of such issues is that ontology extraction has been restricted, so far, to the extraction of a very limited range of relations over limited domains. The Espresso system, for instance, (Pantel and Penachiotti, 2006) returns succession relations between named entities, where quantification is not an issue, and reaction and production relations using as data set an introductory chemistry textbook which, presumably, only contains general statements. We can further cite Ravichandran and Hovy (2002) who extract birthdates, inventions, discoveries and locations for named entities, as well as taxonomic relations of a definitional nature (their extraction patterns, following Hearst, 1992, are extremely specific and may not achieve high recall). Similarly, the KnowItAll system (Etzioni et al, 2004) returns taxonomic relations for named entities such as cities, US states, actors and films. Völker et al (2007), whose system, LExO, produces ontological class descriptions from Wikipedia definitions and a fishery glossary, summarise the issue:

> "our approach is restricted to texts with **definitory character** such as glossary entries or encyclopedic descriptions which have a universal reading and a more or less canonical form [...] In order to extend the applicability of LExO to a greater variety of textual resources, one would need a component for the automatic identification of natural language definitions."

What Völker et al call 'natural language definitions' is close enough to what others have named 'commonsense statements'. The earliest mention of common sense that we are aware of in the AI literature can be found in a paper by McCarthy (1959), where the concept is exemplified as follows:

> "a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows."

Some time later, the idea of 'commonsense knowledge' appeared in the literature:

> "Commonsense knowledge is knowledge about the structure of the external world that is acquired and applied without concentrated effort by any normal

[3]Note that for this task, it is insufficient to restrict the lexical domain of the object as being animals or plants. And even if this was a solution to the problem, we would be faced with the issue of deciding what animals and plants are — in our 2006 paper, we used a manually created resources for this purpose, but this was one of the factors that led to rather poor recall.

> human that allows him or her to meet the everyday demands of the physical, spatial, temporal, and social environment with a reasonable degree of success."
> (Kuipers, 1979)

The idea that obtaining such knowledge might allow AI systems to achieve their goals more easily (planning, conversing, etc.) led to large efforts to construct man-made databases of 'commonsense statements'. The most prominent projects are Cyc (Lenat, 1990) and the OpenMind Commonsense project (Singh, 2002), which respectively employ trained knowledge engineers and web users to obtain ontological knowledge.

The amount of time and effort needed to manually produce such resources being a major barrier to the achievement of AI's aims, some systems have started to appear which attempt to automate the task. Those systems rely on the linguistic concept of **genericity**. We will dedicate most of our second chapter to genericity, so we will simply report here the informal definition given by Cohen (2002):

> "We often express our knowledge about the world in sentences such as the following:
>
> 1. (a) Ravens are black.
>   (b) Tigers have stripes.
>   (c) Mary jogs in the park.
>
> We refer to such sentences as *generics*. They appear to express some sort of generalization: about ravens, about tigers, and about Mary, respectively."

The idea behind the new commonsense extraction systems is that generic sentences like the ones presented by Cohen contain the desired knowledge to create large ontologies of the Cyc type. The problem with this approach, though, is that the concept of genericity is very elusive and it is not entirely clear what sentences should be extracted as 'generic statements'. Suh (2006) and Suh et al (2006) attempt to retrieve commonsense statements from Wikipedia. Their system, however, makes simplifying assumptions with regard to the syntax of genericity: in particular, all bare plurals (and bare plurals only) are considered generic. As we will show in Chapter 2, this assumption is a very weak approximation of the phenomenon. The extracted statements are represented as RDF triples which, Suh admits, are unsuitable to formalise the complex semantics of generics. Dankov et al (2008) also claim to extract generics but their method relies mostly on already quantified sentences (not generic in the classical linguistic sense) and on the frequency of cooccurrence of a particular noun phrase with a particular verb. The former method is sure to provide poor recall (we will show in Chapter 5 that only 7% of all noun phrases are explicitly quantified) while the latter will lack both in recall and precision: there is, as far as we know, no proven correlation between the level of frequency, or even characterisation, of a

verbal predicate with respect to a noun phrase and the notion of genericity. For instance, the Internet query "cats have tails" returns 21,500 matches in Google while the query "cats are vaccinated" produces 62,700 hits, this, despite the fact that the former is an acceptable generic sentence while the latter is not. This is because some verbal phrases are frequent with the existential use of the noun phrase, even though they do not make good characteristics for the concept under consideration.

Given the difficulty of identifying generics using simple syntactic means, it seems that an in-depth corpus study of their behaviour would be needed. We are not aware of any such effort in corpus linguistics but two computational linguistics projects have involved genericity annotations. The ACE corpus (2008) and the GNOME corpus (Poesio, 2000) attempt a broad classification of generic against non-generic entities via human annotations. However, the ACE guidelines do not fully fit the knowledge brought by the linguistic literature with regards to generics and the GNOME corpus is limited to three genres. We give a more detailed overview of both corpora in Chapter 4 of this thesis. Our own attempts at producing genericity annotations guidelines (Herbelot and Copestake, 2008 — see also Section 4.6 of this thesis) showed that the notion is particularly difficult to describe in such a way that both matches the linguistic theory and provides ease of annotation.

Aside from being a complex phenomenon to identify, genericity is not easily formalised. All the attempts we know of in computational linguistics rely on the **default** reading of generics, that is, the observation that generic sentences accept exceptions (for instance, we can utter the sentence *Cats have four legs* even though some of them might lack a limb as a result of an accident or birth defect). These attempts stem from early research on defeasible reasoning: default logic (Reiter, 1980), non-monotonic logic (McDermott and Doyle, 1982; Asher and Morreau, 1991) or again semantic inheritance networks (Horty et al, 1990), the latter being extended by Vogel (1995) to model inference over chains of generic sentences. Vogel (2008) also proposed a model of first-order belief revision which interprets generics in terms of restriction of the denotation set of a certain predicate over a certain domain.

Overall, it is safe to say that computational linguists have mostly avoided the problem of genericity because of its intractability. Cooper et al (1996), in their development of the FraCaS test suite for computational semantics, include generic readings in their coverage of bare plurals but decline to expand on the phenomena producing the observed range of quantification, labelling them as 'poorly understood'.

In this thesis, we will try to overcome the problems caused by genericity by interpreting it in terms of quantification. Our general hypothesis is that quantification is more tractable from the point of view of both annotation and formalisation. The natural language quantifiers, for instance, are well understood by humans and do not need any further explanation when producing annotation guidelines. Further, the assumption that generics

quantify means that their 'quantification value' can be formalised in a semantic system which does not require any complex representation of the domain under consideration —- in contrast to the defeasible reasoning literature.

1.3 Objectives

Having highlighted the motivation for performing quantification resolution, we should spell out the requirements for producing a system able to automate the process. Those requirements will in turn inform the objectives of this work.

We should first point out that quantification resolution, as we have presented it, is a novel task and does not have the support of previously published theories, tools and corpora. We must therefore fulfil basic (but non-trivial) requisites at both theoretical and applied levels.

The practical requirements for the quantification resolution task are standard for all natural language systems. Our implementation should have a wide coverage and be accurate across all classification labels. This high-level requisite implies a need for appropriate training data. A corpus should be available with as many annotated examples as it is possible to produce, and it should cover a large enough range of texts, both topically and stylistically. This, in turn, presupposes the ability to create such a high quality resource with high human agreement, following the theory we will have laid down.

Theoretically, the description of quantification used by the system should conform to two requisites: it should include the linguistic knowledge available on the topic and it should be formalised in a way that our computational goals can be fulfilled. So we should present a view of quantification grounded in the literature on quantification and ambiguity (we will show that the subject of genericity covers a large part of our needs) and we should propose a formalisation that allows such inferences as described in Section 1.1.2.

Our objectives are therefore as follows:

give a formal description of quantification and of the ambiguity phenomenon with respect to quantification

produce annotation guidelines in accordance with the proposed theory, with the objective of creating a quantification corpus

test the annotation guidelines and report annotator agreements on the test set

implement an automatic classification system and train it using data annotated according the proposed guidelines

test the system and report initial results.

1.4 No objects

At this point, we should make clear one limitation of our work; namely, that we will only treat the case of *subject* noun phrases. The main reason for this is that objects differ drastically from subjects, in terms of syntax and semantics. They are also the topic of much disagreement in the linguistic literature. This section justifies our decision to leave them out of this work.

The range of constructs that give rise to genericity in objects is the same as in the case of subjects but their distribution is different. Bare plurals and, only in rare cases of concept reference, (in)definite singulars can express generalisation:

7. Elephants scare lions.

8. In Africa, we filmed the lion. (Gerstner-Link and Krifka, 1993)

It seems at first sight that the observations we made about subject noun phrases are applicable to objects: they can be quantified (at least existentially and universally) and the existential reading allows for inference while the universal reading doesn't:

9. Oscar writes novels = Oscar writes books.

10. John hates lawyers (Cohen and Erteschik-Shir, 2002) $\not\Rightarrow$ John hates people.

However, if there is general agreement in the linguistic literature as to which subject cases can be regarded as existential or universal/generic, this is not the case for objects. Link (1995) proposes, respectively, an existential and generic reading of the following two examples:

11. Cowboys carry guns.

12. Frogs catch flies.

His argument is that, although one can paraphrase 11 with the following:

13. For every (typical) instance x of the kind Cowboy, there is a gun y which x carries.

It is not possible to paraphrase 12 in a similar way:

14. For every (typical) instance x of the kind Frog, there is a fly y which x catches.

However, Cohen and Erteschik-Shir (2002) argue for an existential reading of 12, responding to Link with a new paraphrase based on the assumption that a spatiotemporal variable is introduced by the sentence:

15. In general, if x is a frog and s is a stage, there is a fly y so that x catches y on stage s (sic).

There is also the assumption, in the linguistic literature, that a specific phenomenon blocks the existential reading in some cases of bare plurals. What this phenomenon is is however debated. Cohen and Erteschik-Shir (2002) note that, given the following two sentences, the second one fails to give an existential interpretation:

16. John knows lawyers.

17. John hates lawyers.

Their account is based on a topic/focus distinction. They claim that topics block existential readings and explain 17 by assuming that there is a presupposition to the sentence, *John knows lawyers*, which blocks the existential formalisation in a DRT (Discourse Representation Theory, Kamp, 1981) setting.

Glasbey (2007) disproves this account, though, by claiming that the object in *John hates diamonds* is generic despite the fact that no clear presupposition can be assumed for that statement. Instead, she claims that it is a specific class of verbs — the 'psychological verbs with experiencer subjects', or psych-ES verbs, as defined by Levin (1993) — which fails to give existential readings for bare plurals in object position. It is not clear how her account deals with the following two sentences, which we take as including a generic following a psychological verb with experiencer object:

18. John scares lawyers.

19. Green suits blonde women.

Overall, it seems to be the case that presuppositions have a complex impact on the interpretation of object noun phrases. For instance, formalising 10 as

20. $John'(x) \quad y[lawyer'(y) = \ hate \ '(x, y)]$

is clearly not adequate, as the universal reading of the object implies that John hates all lawyers in the world. This is obviously not true as we cannot assume that John also knows all lawyers in the world — the sentence seems to express some kind of underlying modality whereby John has the potential to hate every single lawyer in the world, but only hates the ones that he actually meets.

As no clear agreement can be found amongst linguists with regard to the quantificational interpretation of object noun phrases, we will leave their study as further work.

1.5 Outline

This thesis is presented as an overview of quantification, from theory to practice. The next two chapters (2 and 3) discuss fundamental theoretical questions in relation to ambiguous quantification. We introduce a basic terminology and relate our own conceptual frame to the existing linguistic literature, then propose a formalisation of the observed ambiguity as underspecification. We then move to more practical matters and discuss the creation of a corpus for quantification. Such an effort requires adequate annotation guidelines. Accordingly, we present in Chapter 4 an annotation scheme mapped onto the theoretical claims made in the first chapters and report the results of annotator agreement experiments. Following this, we turn to the application of our theoretical results by presenting a baseline system for quantification resolution, based on the use of simple syntactic cues (Chapter 5). Having investigated the limitations of our baseline, we argue that quantification resolution relies heavily on pragmatics and that, given enough annotated data, it is possible to use distributional techniques to recover such necessary pragmatic information. Chapter 6 ends with results of experiments showing positive correlations between the performance of the system and what we describe as **situational analogy**. We conclude the thesis with a summary of our main theoretical and applied results, and point at the various questions that are left open, or have been opened, by this work.

Chapter 2

A theoretical account of ambiguously quantified noun phrases

We emphasised in our introduction the importance of resolving quantificational ambiguities for Natural Language Processing. We highlighted in particular the issues experienced when performing information extraction and when reasoning over formal databases or ontologies. The tasks that we considered are fully automated, that is, we make the assumption that the data stored and subsequently processed by a given information system is in machine-readable form. When doing quantification resolution, it is therefore not sufficient to give an explicit, natural language quantifier to those phrases that lack one — the sentence must be appropriately formalised.

In this chapter, we review the issues linked to the formalisation and semantics of ambiguously quantified noun phrases. Having first clarified our terminology, we will turn to the most commonly discussed case of ambiguous quantification, the bare plural, which is known to express both a standard existential reading and a 'generalisation' reading, the latter produced by what linguists have dubbed the **genericity** phenomenon.

Genericity has been extensively studied, especially with regards to its own ambiguity effects. As we will show, it is not clear what 'generalisation' means when it comes to estimating the cardinality of a reference set. The phenomenon also has the interest of occurring in a wide range of grammatical constructs including not only the bare plural, but also definite and indefinite singulars and mass terms, that is, in (nearly) all those constituents that exhibit ambiguous quantification.

A large part of this chapter is dedicated to an overview of the linguistic work devoted to genericity. By presenting the issues experienced when trying to offer a unified semantic and logical account of generic sentences, we will expand on the more theoretical objective suggested in Section 1.3 of our introduction, namely, how we should represent quantification in a formal system. The discussion will open the door to the formalisation of ambiguous quantification that we then propose in the following chapter.

2.1 Terminology

2.1.1 Quantification resolution

We can informally define our task as the translation of **ambiguously quantified** noun phrases into unambiguous ones, this translation being performed by either adding an appropriate determiner to the noun phrase, or by replacing an existing one:

21. Cats are mammals = **All** cats are mammals.
22. Cats were sleeping by the fire = **Some** cats were sleeping by the fire.
23. The Orioles wear white shirts at home = **All** Orioles wear white shirts at home.
24. The beans spilt out of the bag = **Most/all** of the beans spilt out of the bag.
25. Water was dripping through the ceiling = **Some** water was dripping through the ceiling.

More formally, we will talk of our task as **quantification resolution**, that is, the process of taking an ambiguously quantified noun phrase (NP) and giving it a formalisation appropriate to the semantics of the sentence in which it appears. This formalisation should express a *unique* set relation:

26. All cats sleep.

    $\phi \quad \psi \quad = \quad \phi$ where ϕ is the set of all cats and ψ the set of all things sleeping.

In order to fully specify what we mean by 'unique set relation', let us consider the behaviour of various constructions. Determiners such as *a* or *the* — which typically introduce several possible readings for an NP — behave in a similar way to the 'true' quantifiers as far as compositional semantics is concerned. It is convenient to use generalized quantifier notation uniformly. For instance, we can write:

27. $a'(x,\ cat'(x),\ sleep'(x))$[1]

as we would write:

28. $some'(x,\ cat'(x),\ sleep'(x))$

[1]Throughout this thesis, we will be using the prime notation informally to denote the semantics associated with a given word. So *some* is a lexical item while *some'* is the semantics associated with *some*.

The second and third arguments of the determiners can be taken as the two sets that we are trying to formally relate.

However, the quantification semantics of *some* can be fully defined (given a singular count noun phrase, we are talking of one entity only) while that of *a* cannot: in a singular noun phrase introduced by *a*, the referent can either be a single entity or a plurality with various possible quantificational interpretations (contrast *A cat is a mammal* with *A duck lays eggs* — the former is a universal, the latter is not). So we are able to give a shallow logical representation of both determiners, but a formal representation in terms of sets is not available for *a* without further ambiguity resolution.

Note that we are only talking here of **quantification semantics**, that is, of the quantities that the determiner selects in the set under consideration. The full semantics of *a* would require further contextual information. For instance, we would have to ascertain whether the cat in 27 is a specific cat or any cat. The following two sentences give appropriate contexts for the two readings:

29. Mary has a cat (named Tom).

30. Mary wants a cat (any cat).

Further, what we call quantification resolution is not cardinalisation: the aim is not to find out how many cat entities are implied in the sentence *Some cats sleep*, but to find one unique formalisation which accounts for all the possible worlds entailed by that sentence. So whether we are talking about 2, 10, or 100 sleeping cats, we can write:

31. $0 < \quad \phi \quad \psi \quad < \quad \phi$

where ϕ is the set of all cats and ψ the set of all things sleeping. We use the strict interpretation of *some'*, that is, *some cats* can never be *all cats*. We will justify this in Section 2.1.2 as we introduce the logical form of *most'* as an even stricter upper bound for *some'*.

Again, finding a unique formalisation is not possible in the case of an ambiguously quantified noun phrase, like the bare plural *cats* in the sentence *Cats sleep*, where two or more formalisations may clash:

32. $0 < \quad \phi \quad \psi \quad < \quad \phi$ (some cats sleep)

33. $\quad \phi \quad \psi \quad = \quad \phi$ (all cats are known to sleep)

We can thus define our goal further by saying that quantification resolution consists in annotating an ambiguously quantified noun phrase with a fully specified quantifier, and

that a fully specified quantifier is a quantifier for which we have a quantification semantics (as opposed to full semantics) with a unique, unambiguous set relation.

Chapter 3 of this thesis is concerned with finding a representation for quantifiers that incorporates a set relation as described in this section and that is semantically well-motivated. Our goal is to achieve a formalisation of quantifiers which follows from previous work in linguistics and is, at least partially, implementable and usable as logical representation in inference systems. The formalism that we choose, based on the work of Link (1983) —see Section 3.1 — includes a complex representation of plurality. This formalism has previously received partial implementations, as in Copestake (1989). Reduced to its simplest form, it can also be taken as an expression of the natural language quantifiers *one*, *some*, *most* and *all* and be fed to existing inference systems that deal with already quantified statements (see, for instance, the natural logic of MacCartney and Manning, 2008). Although a full implementation would be desirable, our experimental work in this thesis focuses on the parts of the formalisation that are directly translatable to such systems. As such, the set relation expressing the fully specified quantifier in the sentence is of primary concern.

2.1.2 Quantifiers

Natural language quantifiers have traditionally been categorised as either type <1,1> or type <1> quantifiers (Peters and Westerståhl, 2006). Quantifiers of type <1> are properties of sets and are expressed through pronouns like *nothing*, *everybody* or *no one*. They combine with a verb phrase (the **scope** of the quantifier) to form a sentence:

34. Everybody enjoyed the party.

Quantifiers of type <1,1> are binary relations between sets and are expressed through determiners like *some*, *all* or *no*. They combine with a noun phrase (the **restriction** of the quantifier) and a verb phrase (its **scope**) to form a sentence:

35. All guests enjoyed the party.

Note that the subject noun phrase in 35 is itself a type <1> quantifier.

In this work, we will not consider the pronominal type <1> quantifiers, which are — in English at least — semantically unambiguous. Instead, we will turn our attention to the type <1,1> determiners. In line with our definition of quantification resolution, we will posit that to be called a **fully specified quantifier**, a determiner must have *one and exactly one* formalisation. That is, a quantifier unambiguously denotes a set relation. The following are formalisations for *some'*, *most'* and *all'* (the lower bounds for *some'* and *most'* are borrowed from Leslie, 2007):

36. if $some'(\phi, \psi)$ then $0 < \quad \phi \quad \psi \quad < \quad \phi \quad \psi$

37. if $most'(\phi, \psi)$ then $\quad \phi \quad \psi \qquad \phi \quad \psi \quad < \quad \phi$

38. if $all'(\phi, \psi)$ then $\quad \phi \quad \psi \quad = \quad \phi$

We should stress that these set relations assume a division of the semantic space of quantification with no overlap: the formalisations are mutually exclusive. This view is incompatible with a strictly logical interpretation of the quantifiers under consideration, which dictates that if $most'(\phi, \psi)$ then $some'(\phi, \psi)$ and if $all'(\phi, \psi)$ then $most'(\phi, \psi)$. It would be possible to adopt a more traditional formalisation where *some'* and *most'* lack an upper bound without any major consequences for the rest of this thesis. However, we will use 36 to 38 for reasons related to pragmatics. In Chapter 4, which presents a task where humans are required to perform quantification resolution, we assume the Gricean maxim of quantity (Grice, 1975) and expect annotators to choose the most informative quantifier when interpreting ambiguously quantified statements. Under this assumption, *some'* never means *all'*. Formalisations 36 to 38 reflect this pragmatic fact.

Using our single formalisation constraint, we can say that the determiners in 39 are fully specified, while those in 40 are ambiguous quantified.

39. some, most, all

40. a, the

It must be noted that *many'* and *few'* have their own, special behaviour. Much has been said on their semantic ambiguity (e.g. Lappin, 2000). As an example of the issues they cause to the semanticist, we will now expand on their so-called proportional and relative readings, as defined by Cohen (2001).

41. Many Frenchmen smoke.

42. Many Frenchmen eat horsemeat.

Following Cohen (2001), sentence 41 tells us that a high proportion of Frenchmen smoke, while 42 implies a proportion of Frenchmen which is significantly high, compared to, say, the proportion of Britons or Germans that eat horsemeat (but not necessarily high within the set of all Frenchmen). The respective formalisations for *many'* are:

43. $many'(\phi, \psi)$ is true iff $\rho < \frac{|\phi \cap \psi|}{|\phi|} < 1$, where ρ is 'large'

44. $many'(\phi, \psi)$ is true iff $\rho < \frac{|\phi \cap \psi|}{|\cup ALT(\phi) \cap \psi|} < 1$, where ρ is 'large' and ALT is the set of alternatives to ϕ[2]

[2]We have added an upper bound to Cohen's formalisation under the assumption that *many'* is never *all'*.

So getting one unique formalisation for *many'* actually requires some sort of disambiguation if we want to get to the full quantification semantics of the noun phrase. This would be a reason to talk of ambiguous quantification. However, there is a fundamental difference between *many'* and, say, *a'* or *the'*. Under the assumption of our tripartite system, where *some'*, *most'* and *all'* share the quantificational space, the formalisations available for the *a'* and *the'* are not reducible to a common denominator (other than the whole quantificational space), while the formalisations for *many'* can be shown to cover the same area of that space. It should be clear that, as long as ϕ is not empty, both 43 and 44 entail:

45. $0 < \quad \phi \quad \psi \quad < \quad \phi$

which happens to cover the quantificational space of *some'* and *most'*, as we defined it in 36 and 37[3]. We can thus say that *many'* is consistently quantified — the two readings cover the same space. Less formally, we can also say that the two readings are in a sense 'compatible': 41 could probably be shown to be true under both formalisations. Lappin (2000) talks of 'underspecification' in *many'* and *few'*. We will reserve the term for a different phenomenon, but the idea should be clear: there is quantificational consistency across all readings of *many* and *few*.

We will also remark that if various bare plurals can be paraphrased using *some*, *most* and *all*, as well as the relative reading of *many* (see next section), they don't seem to be paraphrasable by the proportional reading of *many* or by *few*. In fact, as we will argue later in Section 3.2.1, *few'* has a negative polarity which seems incompatible with the semantics of bare plurals — even when the quantification of the noun phrase under consideration implies a small number of individuals:

46. Mosquitoes carry malaria = Few mosquitoes carry malaria

We will claim at several points of this thesis that *relatively many* is a good paraphrase for a certain class of bare plurals, but we do not believe that the quantification implied by *relatively many* is in any way different from that of *some*. Therefore, we will overall regard *many'* and *few'* as simple existential forms.

2.1.3 Scoping the study

We are interested in three grammatical forms which give rise to quantification ambiguity in noun phrases: the definite form, the indefinite singular *a* and the bare form, that is

[3]This is not to say that *many'* is ambiguous between *some'* and *most'*: it is possible to utter, with the same intention, *Many Frenchmen smoke* whether the actual proportion of smoking Frenchmen is 40% or 80%.

the absence of an explicit determiner, as in bare plurals and mass terms. We give below examples of those constructs, together with appropriate, quantified readings for each of them.

Definites

47. The cat is a mammal. (The kind *Cat*/ All cats).

48. The cat can sleep rolled up, with its head on its hind legs. (The kind *Cat*/ Most cats — those without arthritis/ That particular cat).

49. The cat was sleeping by the fire. (That particular cat).

50. The Galapagos turtle lives over 150 years. (The kind *Galapagos turtle*/ Some lucky Galapagos turtles).

51. The dodo is extinct. (The kind *Dodo*).

52. At the end of the lecture, the/her/his/their students asked questions about the dodo. (Some of the/Some of her/Some of his/Some of their students).

The indefinite singular

53. A cat is a mammal. (The kind *Cat*/ A stereotypical cat/ All cats).

54. A cat can sleep rolled up, with its head on its hind legs. (The kind *Cat*/ A stereotypical cat/ Most cats — those without arthritis).

55. A cat was sleeping by the fire. (That particular cat).

The bare plural

56. Cats are mammals. (The kind *Cat*/ All cats).

57. Cats can sleep rolled up, with their heads on their hind legs. (The kind *Cat*/ Most cats — those without arthritis).

58. Cats were sleeping by the fire. (Some cats).

59. Galapagos turtles live over 150 years. (The kind *Galapagos turtle*/ Some lucky Galapagos turtles).

60. Dodos are extinct. (The kind *Dodo*).

The bare singular

61. Water is necessary for life. (The kind *Water*).

62. Water was dripping through the ceiling. (Some water).

63. Furniture has a practical purpose. (Most furniture — except contemporary art tables and chairs).

2.1.4 Reference: sets and lattices

We have already referred to quantification resolution in the introduction as the process of quantifying the **referent** of the noun phrase, i.e. the set of entities denoted by the NP. We must now further specify our notion of reference.

In model-theoretic semantics, given a noun phrase, the Nbar of that noun phrase denotes the set of all entities for which the property indicated by its lexical realisation is true. For instance, in the NP *some dogs*, the Nbar denotes the set of all things of which the property *DOG* is true. Further, the NP denotes the set of all sets that are in a certain intersection relation with the Nbar denotation. The exact nature of the intersection is given by the quantifier of the noun phrase. So in *some dogs*, the NP denotes the set of all sets $Q_{1...n}$ which intersect with the Nbar denotation D in a way that $0 < \quad D \quad Q_{1...n} \quad < \quad D \quad Q_{1...n}$.

We take a lattice view of plurals and mass terms (see Link, 1983, and Section 3.1 of this thesis) where any point of the lattice under the supremum refers to a proper subset of the supremum. In this view, the supremum corresponds to the **Nbar referent** while the **NP referent** might point at any other point in the lattice. So in sentence 58, for instance, the Nbar referent can be taken as *all cats in the world* while the NP referent is *those cats sleeping by the fire right now*. It is possible to relate the lattice interpretation to the classical idea of denotation by saying that the supremum is the maximum plurality of entities with property P while the other points in the lattice denote the entities in the intersection between the 'classical' Nbar denotation and NP denotation. There are arguments against associating points in a lattice with the notion of set (Link, 1983; 1998) – in particular the fact that the set representation is not fully adequate for the formalisation of mass terms. However, Landman (1989) refutes Link's arguments and claims that it is possible to give a set theoretic interpretation of lattices. We follow Landman and throughout this thesis, we will use the term 'set' to refer to pluralities.

2.2 From bare plurals to ambiguous quantification: the genericity phenomenon

2.2.1 Bare plurals

Bare plurals are interesting for several reasons. First, they have two accepted readings, one existential (which can be paraphrased using the determiner *some*) and one so-called generic reading. Sentences 64 and 65 are examples of those two readings respectively.

64. Dogs were in my garden yesterday.

65. Dogs make good pets.

Secondly, the generic reading itself is known to have several possible semantic interpretations, some of which cannot be easily paraphrased using simple natural language quantifiers:

66. Turtles are reptiles (all)

67. Turtles live over 100 years (some)

68. Turtles lay eggs (most healthy mature female turtles...)

Lastly, in some sentences, bare plurals are commonly used as paraphrases of definite and indefinite singulars, suggesting that some of the readings associated with bare plurals are available to other ambiguously quantified constructs. (We will see later that those are not exact paraphrases, but the constructs are sufficiently close that the definite and indefinite singulars can also be called generic.)

69. Cats are mammals.

70. A cat is a mammal.

71. The cat is a mammal.

Whether bare plurals can actually be called ambiguous or not has been extensively written about. The kind reference analysis proposed by Carlson (1977) suggests that bare plurals always refer to kinds and that the apparent existential reading observed in some sentences is produced by the presence of an episodic verbal predicate:

72. Dogs were in my garden yesterday.

The ambiguity analysis suggested by Gerstner-Link and Krifka (1993) prefers to see bare plurals as inherently ambiguous between the existential and the generic reading. Further, Krifka (2004) argues that bare NPs are formally properties of individuals rather than a direct reference to those individuals, and are therefore themselves neither kind referring nor ambiguous but can take either an existential or a generic interpretation via type-shifting (we will come back to this interpretation at several points in this chapter).

We will not make any contribution to this argument, as it concerns the nature of the linguistic construct — the kind *bare plural* — taken in isolation. In this work, we are primarily interested in instances of that kind, that is, in phrases such as *cats* or *white elephants*, which point at two (or more) possible reference sets, variously quantified. Therefore we will carry on talking of 'ambiguous bare plurals' as a short form for 'ambiguous bare plural instances'. We will also, at first, follow the existential/generic binary distinction suggested by the literature and assume that the quantifier of a bare plural is either *some* or the so-called *GEN* operator. The formalisation of *some* should by now be clear, so we will expand on what is understood by genericity, and the form that its quantifier, *GEN*, should take.

2.2.2 Genericity: some definitions

In the *Generic Book*, Krifka et al (1995) introduce genericity as two separate phenomena: one focuses on the sentence and is described as a way to 'report regularity'; the other one focuses on the noun phrase itself and is described as **a reference to a kind**. The former is also known as **characteristic predication**.

Characteristic verbal predicates stress the habitual character of an action, as in:

73. John smokes a cigar after dinner. (Habitually).

As for the kind-reference phenomenon, Krifka et al introduce it with the following example:

74. The potato was first cultivated in South America.

The authors point out that the subject noun phrase in this sentence does not designate 'some particular potato or group of potatoes, but rather the kind Potato (Solanum tuberosum) itself'.

In the same vein, Cohen (2002) introduces the concept of genericity with some examples of mixed predicates involving both kind reference and characteristic predication and notes that those examples 'appear to express some sort of generalisation'. Behrens (2005) follows the same type of definition: 'Generic statements express generalizations about kinds. A classic generic sentence contains a kind-referring noun phrase as its topic'. Other such definitions include Khemlani et al (2008) or again Leslie (2008).

Whether habituality and kind-reference should be encompassed by the same terminology is not clear, and some prefer to see the two phenomena as separate (e.g. Heyer, 1990). The main reason for including them under the same umbrella is the fact that some kind-referring sentences seem to implicitly express typicality. So for instance, 75 seems to behave similarly to 73.

75. (typically/normally/habitually) The dog barks.

However, the semantics attached to habituality is not obviously present in examples such as:

76. ??The lion always/often/usually/rarely/never is a species.[4] (Heyer, 1990)

A slightly earlier account of generic constructs, that of Gerstner-Link and Krifka (1993), keeps habituals outside of the classification. It distinguishes instead between so-called D-generics and I-generics, the former being a true reference to kinds while the latter is closely linked to the phenomenon of modal quantification where the rule $A \quad B$ does not translate as *If A then B* but rather as *If A then probably B*. The terminology refers to some rough classification where definite singular NPs belong to the first group, while indefinite singular NPs belong to the second. Examples of non-overlapping generic uses are given by the authors:

77. (The/??A) dodo is extinct.

78. In Kenya they filmed (the/?a) lion.

79. (The/?An) antelope gathers near water holes.

80. (The/?A) rat reached Australia in 1770.

81. (The/?A) madrigal is popular.

82. (?The/A) green bottle has a narrow neck.

We should remark that all examples above could be paraphrased with bare plurals. Note also that when we talk of 'non-overlapping use', we mean that the article cannot be changed without altering the semantics of the sentence. It is actually possible to utter *An antelope gathers near water holes* but the sentence can then only refer to a subkind

[4]Throughout this thesis, we use the double question mark notation ?? to indicate that a sentence is altogether infelicitous. We use a single question mark in cases where the sentence is infelicitous for the meaning under consideration but would be acceptable in another context: see Example 79 and the associated following paragraph.

of the kind *Antelope* (perhaps the Arabian oryx). We won't talk about subkinds in this chapter but we will touch on the topic in Section 3.4.2.

Heyer (1990), who also refuses the inclusion of habituals in genericity, prefers to talk about reference to kinds and defaults respectively. (We will come back to the default reading of generics in Section 2.2.4.)

As our focus is the quantification of the noun phrase as opposed to the semantics of the generic sentence, we will prefer the Gerstner-Link and Krifka classification, which makes obvious reference to the grammatical constructs typical to the two types of genericity. We will however prefer to take those grammatical constructions as the basis of our classification rather than their supposed semantics and we will talk of **D-generics** to refer to noun phrases where an indefinite singular paraphrase is infelicitous, and conversely, of **I-generics** where a definite singular paraphrase is not available. Many NPs, of course, can take either construct, and we will call those **mixed generics**. A stricter classification means that we must also find a space for those generics that are only available in bare form. We will give those a separate class, the **bare generics**:

83. Dodos are extinct. D-generic

84. Birds lay eggs. I-generic

85. Ducks lay eggs. Mixed generic

86. Cars have radios. Bare generic[5]

87. Water is necessary for life. Bare generic

The concept of genericity introduces difficulties in terminology. In particular, it is not clear what is understood by **kind**. Example 74 implies that the authors of the *Generic Book* see the concept as close to that of species. Krifka seems to later alter this idea, though, by writing that 'kind reference involves reference to an entity that is related to specimens' (2004) — implying an idea of stereotype more than species.

There are fortunately more formal definitions in the literature. According to Carlson (1977) kinds, together with 'objects' and 'stages', make up a basic ontology where objects can 'realise' kinds and stages can 'realise' objects or kinds. So when we are saying that *John smokes after dinner*, we are saying that John is a stage of a kind (the kind *Human*, or maybe *John*), and the part of John that smokes every evening after dinner is a stage of John. This conceptualisation presents kinds as abstract entities linked to concrete entities (their instances) via the process of realisation. The theory, however, doesn't make any claims about which abstractions can be regarded as kinds.

[5]We argue that the sentence *A car has a radio* differs semantically from *Cars have radios*. In Section 3.2.2 we mention a psychological experiment (Leslie et al, 2009) which shows that a majority statement such as 86 is considered unnatural by human subjects when converted to an indefinite singular generic.

Another, philosophically different, view of kinds is that they do not refer to any abstraction but are rather the collection of their instances. So we can say that a kind refers to concrete entities and therefore spans multiple points in space and time (Gerstner-Link and Krifka, 1993; Chierchia, 1998).

One point of agreement amongst theorists seems to be that some grammatical constructs are not conducive to kinds. For instance, in the following, *his beans* would not normally be regarded as a generic noun phrase, even though there seems to be some generalisation applied to the beans sold on the market:

88. I always buy my vegetables from the man on the market. Everything on his stand is locally produced and *his beans* are delicious.

There seems to be some implicit understanding that small sets do not make kinds. For instance, the set of chairs in Mary's lounge does not relate to any recognisable kind such as *Chair-in-Mary's-lounge*. Dahl (1975) gives some further linguistic foundation to this claim by remarking that generic bare plurals cannot be applied to restricted sets of a more general class. The following example is taken from Gerstner-Link and Krifka (1993):

89. There were lions and tigers in the circus ring.

    (a) (Every lion)/(each lion)/(most lions)/roared.

    (b) ?Lions roared.

It could be argued, however, that the problem here is not the application of genericity but the lack of definiteness where expected: replacing 89b with *The lions roared* makes it completely acceptable. What seems more telling is that the sentence *Lions pacing in the circus ring roared*, taken in isolation, cannot be read generically, showing that there are indeed concepts that are more suited to a kind reading than others (compare with *Dinosaurs roared*). Krifka et al (1995) suggest that kind readings only occur in sentences referring to **well-established** kinds, their justification being that, while 90 is possible, sentence 91 is odd.

90. The Coke bottle has a narrow neck

91. ?The green bottle has a narrow neck.

There are however 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):

92. The newly-hatched fly is a lazy insect.

93. The well-crafted bottle has a narrow neck.

While we will show later that sentences such as 88 have more to do with genericity than previously assumed, we regard Dahl's argument as a sign that generic bare plurals can indeed only denote a limited set of concepts.

Given the difficulties encountered when trying to define the concept of kind, we will simply make the assumption, for the rest of this chapter, that genericity occurs in sentences that refer to concepts at different levels of abstraction (the definite singular seems to fit the elusive notion of kind better than bare plurals, which are more centred on instances). We will also refer to I-generics, D-generics, bare generics and mixed constructs as the four ways to express genericity.

2.2.3 The GEN operator

The logical form of generics is the object of much debate. Most accounts agree on the use of a GEN operator in I-generics (Gerstner-Link and Krifka, 1993):

94. $GEN\ x_1 \ldots x_n; y_1 \ldots y_n[Restrictor(x \ldots x_n); Matrix(x_1 \ldots x_n, y_1 \ldots y_n)]$

Roughly, the operator has a traditional quantifier form, where the variables in the restrictor are bound by the quantifier and those in the matrix are bound existentially with scope only in that matrix. Contrarily to other quantifiers, though, the GEN operator does not have an explicit, pronounced form in any known language (see Krifka et al, 1995). This makes the task of semantic interpretation particularly difficult.

It is usually accepted that generics differ from universals in that they accept exceptions: *Lions have four legs* is true, even though some lions might have three legs as a result of a birth defect or accident. Some generics, however, are clearly puzzling in that the number of exceptions seem to outweigh the number of regular entities: *Turtles live over 100 years.*

As discussed in the introduction, the main problem with a formalisation relying on a unique quantifier is that it makes it impossible to perform simple logical proofs on its propositions. Consider the following, which assumes universal quantification on generics:

95. Turtles are reptiles.

    $x; turtle'(x) \quad reptile\ '(x)$

    $turtle'(Tim) \quad reptile\ '(Tim)$ (If Tim is a turtle, Tim is a reptile)

96. Turtles live over 100 years.

    $x; turtle'(x) \quad live100Years\ '(x)$

    $turtle'(Tim) \quad live100Years\ '(Tim)$ (If Tim is a turtle, Tim will live over 100 years)

The logical entailment is true in the first example but false in the second (it is actually unlikely that Tim will make it to his 100th birthday). In fact, it is very difficult to predict the extension of the generic NP without possessing appropriate world knowledge, as demonstrated by the following:

97. Turtles are reptiles (all turtles are reptiles).

98. Turtles lay eggs (healthy, female, mature turtles lay eggs).

99. Turtles live over 100 years (in exceptional circumstances, turtles live over 100 years).

Cohen (2002) gives a good overview of the issues involved, showing that a traditional quantification theory cannot account for the wide variety of generics observed in natural language. The problem leads Gerstner-Link and Krifka (1993) to a questionable move where they argue that the GEN operator can be interpreted in terms of non-monotonic inference rules (something similar to defaults — we will introduce the relevant semantics in the next section) in all cases where 'there should be *no reason why the matrix does not hold*'. It is not clear how the semanticist should treat the examples left aside: for instance, it is difficult to account for 99 with GEN.

To add to the formalisation problem, previous authors have noted that anaphoric sentences worsen the issue. Consider the following, from Krifka (2004):

100. Watermelons contain iron, so John often buys them.

Krifka notes that while the first instance of the noun *watermelons* has a kind reading, its anaphora has an existential reading. It is not clear how this is possible, since the anaphoric reference should point at the same entity as the subject of the sentence. Krifka proposes that bare plurals, as such, do not refer to either kinds or existentials but essentially to properties and take one or the other interpretation depending on context (see also Chierchia, 1998). A formalisation for such a theory is possible but involves complex type-shifting.

Given the problems encountered when trying to formalise GEN, we could ask, like Cohen (2002), if generics do ever quantify. We could, for instance, assume a rule-based theory (see Carlson, 1995) where generic sentences do not say anything about individual entities but express a general rule of the world, whether that rule is biological, sociological, or relates to any other convention. (There are more such interpretations, which give various amounts of importance to individuals, and we will go through some of them in the next section). Or again, we could simply assume that generics are similar to proper nouns (see Carlson, 1977) and give generic sentences a simple subject/predicate structure. This is, after all, the simplest possible formalisation for some D-generics:

101. The dodo is extinct: $extinct'(Dodo)$

It is however impossible to abandon quantification for simple pragmatic reasons. It should be clear that if we announce to someone that *Unicorns are mammals*, and subsequently ask them to tell us whether a particular unicorn is a mammal, they will have no difficulty in answering positively. Despite potential differences in determiner semantics, the scenario will be repeated if the initial sentence becomes *The unicorn is a mammal* or *A unicorn is a mammal*. So, unless we are faced with a true kind statement such as 101, we should be able to reason with the formalisation of a generic sentence so as to make inferences about class instances, regardless of the grammatical form of the generic construction. Attributing a subject/predicate structure to all generics does not take into account the way that humans normally reason with such structures. In fact, we will argue that it is even questionable to give a subject/predicate formalisation to all D-generics, as the following example shows.

102. The/?An antelope gathers near water holes. (Gerstner-Link and Krifka, 1993)

Although sentence 102 is a true D-generic, treating the subject like a proper noun ignores the fact that the predicate must collectively apply to individuals in order to be true. The point can be illustrated through an example of anaphora:

103. Mary didn't know that the antelope gathers near water holes, so she was amazed to see a whole group of them this morning by the lake.

It is possible to again assume a type-shifting solution like that of Chierchia (1998) and Krifka (2004) and read *antelope* as a property. This is however actually a sign that instances should be taken into account in the formalisation: indeed, the type-shifting solution implies that a kind is a function that returns the maximum plurality of objects that have a certain property, i.e. it refers to individuals.

A similar argument can be made with regard to bare generics:

104. Mosquitoes carry malaria.

105. Water consists of H_2O.

Mosquitoes can be said to carry malaria in virtue of some individual mosquitoes carrying the disease, and it can be inferred from 105 that individual 'instances' of water consist of H_2O. The view that mass terms have instances in the form of non-overlapping parts fits in with the linguistic hypothesis that they are related to plurals (see for instance Chierchia, 1998). We will come back later to the linguistic notion of 'plural' mass terms (see Section

3.2.4). For the minute, we will just give an idea of what this means in ontological terms. Considering mass terms as instantiable means that any quantity of water is a part of water, like any sofa can be a part of furniture or any message a part of information. This assumption raises philosophical questions that have long been debated. For instance, how small can a part of water be and still be water? Water is still water at the molecular level but not if one starts breaking down the molecules into separate atoms of hydrogen and oxygen. This was discussed at length by Quine (1960). We will not make any contribution to this argument but just assume that humans can select the appropriate referent when thinking of a part of something. This hypothesis is not unfounded as Sharifian and Lotfi (2003) show, for example, that Persian allows for nouns to be treated as mass or count terms depending on the referent that the speaker has in mind: using a plural form for *sugar*, for instance, indicates the denotation of individual granules of sugar.

We will therefore take the stance that generics 'ambiguously quantify' via their whole range of grammatical constructs, from definite to indefinite singulars, and from bare plurals to bare singulars. This has the effect that, for bare plurals, the general idea of ambiguous quantification supersedes the usual existential/generic binary: we take *some* as one of the possible quantifications for the construct.

2.2.4 Semantics of the generic sentence

There have been many semantic theories proposed to account for the seemingly intractable generics. All of them run into problems, the most pervasive one being that it seems practically impossible to account for all possible generic forms found in natural language. In this section, we present the proposals that can be directly linked to particular formalisations and discuss their coverage and limitations. Note, though, that we do not claim to give a comprehensive summary of the semantics of generics. A full account is given in Cohen (2002).

Rules and regulations theory

We have already alluded to Carlson's rules and regulations theory in the last section. Its main idea is that generic sentences have a subject/predicate form, logically expressed by the function $\psi(\phi)$, where the predicate ψ indicates a property of a kind ϕ. Besides the pragmatics-related issues that we identified in Section 2.2.3, the proposal doesn't make any attempt to clarify which properties are acceptable for a given kind. One issue often mentioned is why, for instance, the sentence *Turtles are female* is infelicitous. Leslie (2008) actually gives a solution to this problem, which relies on the presence or absence of a positive alternative property: it is possible to say that *Turtles lay eggs* as there is no alternative property to laying eggs (only the negative property of *not laying eggs*) but it is infelicitous to say that *Turtles are female* as there is one, positive, alternative property,

i.e. the property of being male. Despite this solution, the rules and regulations theory hits other issues when considering the problem of scope ambiguity. In particular, it cannot account for the two readings of sentences like 106.

106. Storks have a favourite nesting area. (Schubert and Pelletier, 1987)

One reading assumes that every stork has a favourite nesting area of its own, while the other reading implies that many storks choose to nest in a particular, identifiable area. Because the distributive reading is not available in a simple subject/predicate logic form, one must assume that a unique nesting area is referred to. This makes the rules and regulations approach rather unattractive as a unified theory.

Inductivist approaches

Inductivist theories approach the semantics of generics through quantification. Their main areas of investigation are the choice of an appropriate, single quantifier to account for all possible instances of genericity, and the choice of the individuals to be quantified over. That is, there is an idea that some individuals may or may not be relevant to the quantification.

We have already shown that it is difficult to pick a single quantifier that covers all cases of observed genericity. To solve this issue, some papers assume universal quantification and focus on restricting the set of relevant individuals to give the correct cardinality. Krifka (1995), or again Pelletier and Asher (1997), offer a version of genericity semantics where the universal quantifier is applied to all 'normal' individuals. It seems, however, complex to define 'normal', and the theory requires an additional domain-restriction approach to deal with sentences such as *Turtles live over 100 years*. Similar problems occur when trying to define the referent of a generic noun phrase using the concepts of 'prototype' or 'stereotype' (see Krifka et al, 1995).

Close to the idea of typicality is that of defaults. Krifka (1987) and Heyer (1990) argue that default logic can be used to formalise generics:

107. Turtles have four legs

$$\frac{Turtle'(x) : HasFourLegs'(x)}{HasFourLegs'(x)}$$

(Unless contrary evidence is supplied, Tim the turtle has four legs).

It is however not obvious how the approach can be expanded beyond simply quantified statements. The next two examples are rather dubious ways to convert more complex generics into defaults. The first involves inserting a negation in the formalisation:

108. Turtles live over 100 years

$$\frac{Turtle'(x):\neg Live100Years'(x)}{\neg Live100Years'(x)}$$

(Unless contrary evidence is supplied, Tim the turtle will not live over 100 years.)

The second involves spelling out complex world knowledge:

109. Turtles lay eggs.

$$\frac{Turtle'(x):Female'(x)\wedge Adult'(x)\wedge Fertile'(x)}{Lay{-}Eggs'(x)}$$

(Unless contrary evidence is supplied, and as long as Tara the turtle is female, adult, and fertile, it will lay eggs.)

In a slightly different vein, Cohen (1996) suggests that generics express probabilities, that is, given a generic noun phrase ϕ, an instance of ϕ has probability P to take part in the predicate ψ of the generic sentence: or in terms of sets, $\frac{|\phi\cap\psi|}{|\phi|} > P$. It is however unclear which value P should take (Cohen suggests 0.5) and whether it should be a constant. We should for instance be able to account for the fact that it is far more likely for a given cat to be a mammal than for a given turtle to lay eggs.

Leslie's three-fold classification

Stepping away from the quantificational puzzle, Leslie (2008) attempts to give an account of the semantic variety of generics by classifying generic statements into either **characteristic**, **majority** or **striking** statements. The characteristic statements, according to her, express an essential property, or 'characteristic dimension' of a kind. For instance, she argues, the characteristic dimensions for an animal species are their mode of reproduction, their diet and their habitat. Striking statements express a noticeable fact or peculiarity about a kind. Majority statements, as their name indicates, just express a statistical fact:

110. characteristic: ducks lay eggs.

111. majority: cars have radios.

112. striking: mosquitoes carry malaria.

Although Leslie doesn't offer any formalisation for her theory, it is fairly easy to map her classes to previously discussed logical forms. The striking class seems to be a semantic interpretation of the *relativelyMany'* quantifier introduced by Cohen (2001) — which we will regard as *some'* (see Section 2.1.2). The majority class naturally maps onto *most'* and the characteristic class can be expressed in the rules and regulations manner, via

a simple subject/predicate form. The obvious issues with the proposal are those that were expressed with regard to the rules and regulations theory. Further, the partitioning introduces difficulties as to what an essential property should be, and by extension, where the boundary between characteristic and striking actually is. According to Leslie, a generic statement is characteristic if it fills a 'characteristic dimension' of a concept. Characteristic dimensions are things like diets or modes of reproduction for animal species and function or role for artefacts or social kinds. Leslie argues that such dimensions, whether innate or the product of early nurture, guide language acquisition: knowing that cats eat meat and seeing her first grazing cow, the child might deduce that *Cows eat grass*, i.e. she specifically looks for information that will fill the 'diet' dimension in her concept of 'cow'.

It is not quite clear, then, why Leslie classifies statements such as *Lions have manes* or *Bees gather honey* in the characteristic class. We could of course imagine characteristic dimensions such as 'striking physical feature' or 'main occupation' but given that many animals lack fillers for those dimensions, it is difficult to accept them as guides for language acquisition.

Despite the terminology issues, Leslie's theory has been so far little criticised and has stood up to psychological experiments involving her classification (see Section 2.2.6 and, in the next chapter, Section 3.2.2). We will therefore return to it in our own account of quantification.

2.2.5 Interaction with other linguistic phenomena

We will finish our overview of the genericity phenomenon by briefly pointing out some important interactions between genericity and other linguistic constructs.

D-generics and collectives

It can be shown that generic noun phrases occurring in the context of a collective predicate have affinities with D-generics: it is usually not possible, looking at a collective generic form, to make inferences at the instance level. So for instance, when we say *Humans generate a phenomenal amount of waste* we can't deduce that given Sandy, a human instance, Sandy generates a phenomenal amount of waste. Those generics are therefore very close to pure kind noun phrases such as *the dodo* in *The dodo is extinct*. This remark will become important when we formalise kinds in Section 3.4.2. We also note that the mass semantics effect observed in such sentences is not covered by more complex accounts of genericity such as that proposed by Leslie.

I-generics and specificity

Krifka et al (1995) comment on the strong interaction between non-specific noun phrases and what they call characterising sentences. Translating this to our terminology, we will simply say that there is a link between non-specificity and indefinite noun phrases (expressed via I-generics and mixed generics). We will first show this informally, and then come back to the definition of specificity.

Note that in the following examples, the subject noun phrase can be translated into a bare plural and express, as appropriate, what Leslie would call characterising or majority statements. However, the use of the indefinite article also indicates as potential referent a single entity, the identity of which is not known. That single entity could be described as non-specific.

113. A cat is a mammal = given a random cat, that cat will be a mammal.

114. A cat has four legs = given a random cat, that cat will probably have four legs.

This aspect of the semantics of indefinite noun phrases reinforces the default interpretation of Krifka (1987) and Heyer (1990). However, the presence of non-specificity is not entirely clear in the class of statements that Leslie labelled as 'striking'.

115. A Frenchman eats horsemeat.

When such a sentence is uttered, it is rather in the context of specifying what the typical Frenchman should do to call himself a Frenchman (we will come back to such normative effects in Section 3.2.3). The semantics here is closer to the idea of prototypes and stereotypes as to that of default. It is therefore debatable whether a sentence such as 115 contains any non-specificity, or whether it refers to one, single, instance: the typical Frenchman. The extent of the relation between indefinites and non-specificity is therefore a bit more constrained than we originally suggested.

Having roughly formulated the issue surrounding the relation between genericics and non-specifics, we must say something about the formal definition of specificity. As argued by Jørgensen (2000), it is not a well-defined concept. The idea behind the notion is that specific entities are identifiable while non-specific ones are not. Jørgensen, 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." (p. 15)

Jørgensen 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 applies. The latter cases are non-specific. In the rest of this work, this is the definition we shall adopt.

2.2.6 Genericity in psychology

The genericity phenomenon has only recently started to be investigated in psychology. The results of early experiments are partly puzzling. For instance, studies have shown that the intractable generics are acquired earlier by children than quantifiers such as *some* or *all*. Hollander et al (2002) show that three-year-olds are able to answer generic questions such as *Are fires hot?* or *Do fish have branches?* while they are unable to answer the same questions when they contain quantifiers (*Do some fish have branches?*, *Are all fires hot?*)

Additional experiments contribute to the idea that, even in later life, humans do not find it easy to correctly map generics to an appropriate quantifier, and vice versa. Khemlani et al (2007) showed that when asked to provide the truth value of a characteristic generic quantified with *all*, human subjects wrongly agree with the statements:

116. All turtles lay eggs.

117. All elephants have tusks.

Humans don't make the mistake for statements simply expressing a majority (*All cars have radios*) or a striking statement (*All turtles live over 100 years*).[6] Khemlani's hypothesis with regard to this mistake is that, given a complex question (and this is a complex question because the set of all egg-laying turtles is not so easily computed), humans fall back on the best possible approximation of the statement — and the less taxing — which is in this case the equivalent bare plural.

Khemlani et al (2008) reported, in a further experiment on syllogisms, that humans also make the reverse mistake. Subjects were presented with questions such as the following:

118. Kangaroos are polymorphic

    Polymorphic individuals have gene Gamma-64

    What follows?

[6] We follow here Khemlani's use of Leslie's terminology.

The sentences used were semantically empty so that the participants could not apply their world knowledge when providing their answers. Without world knowledge, all can be said about a generic sentence is that it existentially quantifies. So the answer to a syllogism such as 118 is *Nothing follows*. Most respondents, though, incorrectly quantified the subject noun phrases as universals, providing answers such as *Kangaroos have gene Gamma-64*.

Leslie (2008) argues accordingly that generics are a default construct of language, used when quantificational processing fails or is too taxing for the hearer.

2.3 Definite plurals: in or out?

We should by now have shown that the phenomenon of genericity covers most of the so-called ambiguously quantified constructs that form the object of this thesis. One has been left aside, though: the definite plural. We have previously commented (see Section 2.2.2) that definite plurals are not traditionally included in the range of grammatical constructions related to genericity. The reasons for this are fairly clear. First, they cannot substitute for other constructs which are typical of genericity:

119. The cat is a mammal.

120. A cat is a mammal.

121. Cats are mammals.

122. ?The cats are mammals.[7]

Further, they tend to appear in contexts where reference is made to instances (i.e. to a subset of a kind) rather than to the kind itself:

123. The cats have fallen asleep by the fire. (My cats, the cats in my household.)

Finally, definite plurals are usually seen as universals. Lyons (1999) argues the point by giving the following example:

124. - I have washed the dishes.

- No you haven't. You have only washed half of them.

[7]Sentence 122 is actually felicitous under a subkind reading (i.e. *All species of cats — the lion, the tiger, the leopard, etc — are mammals*). This reading, however, is not equivalent to that of Sentences 119 to 121, which clearly include individual cats in their semantics.

However, as we briefly showed in Section 2.1.3, the universal interpretation is debatable:

125. The students asked questions about the dodo. (Some students in the relevant set).

We will argue for now that if definite plurals do not partake in the genericity phenomenon, they share similarities with it when it comes to their behaviour towards quantification. At any rate, in the next chapter we will consider them under the same microscope as the traditional generic constructs.

Chapter 3

Underspecified quantification

We introduced several questions in the last chapter. What is a kind? Given a generic concept, how can we reason over the instances of that concept? How should we account for the various semantics of generic noun phrases? Are definite plurals really outside of the genericity phenomenon? In this chapter, we will propose an approach to the formalisation of ambiguous quantifiers which answers those questions while taking care of effects observed in side phenomena such as anaphora and collective constructs.

Having first introduced the notation in use in the rest of this thesis, we will suggest that the silent *GEN* operator hypothesised by most genericity theorists is simply an underspecification effect. Secondly, we will argue that as far as quantification is concerned, definite plurals fall under the same phenomenon. We will then show that the quantification of an NP referent can be formalised as a partitive relation between the Nbar set and the NP set and that this form also allows for correct anaphora resolution in complex cases involving kinds.

3.1 Link's notation (1983)

As the later parts of this chapter contain a fair amount of formalisation, we now introduce the notation that we will be using. The background assumption for our formalisations is that, following Link (1983), plurals can be represented as lattices (see also Section 2.1.4 on reference). A lattice is a partially ordered set in which any two elements have a unique least upper bound (their **join**) and a unique greatest lower bound (their **meet**). The lattices described by Link are join-semilattices, i.e. only the join constraint is enforced.

In what follows, we define each item of notation used in this work, as borrowed from Link. To make things clear, we illustrate the main points via examples over a closed world W containing three cats (Kitty, Sylvester and Bagpuss).

The star sign * generates all individual sums of members of the extension of predicate P. So if P is *cat'*, the extension of $^*P$ is a join-semilattice representing all possible sums of

cats in the world under consideration. The join-semilattice of cats in world W is shown in Figure 3.1. is the individual sum sign; in lattice terms, it is the join operation of the lattice. Note that within Link's theory, a single join operator cannot be used as the join of two count individuals and that of two parts of matter are semantically different (see below).

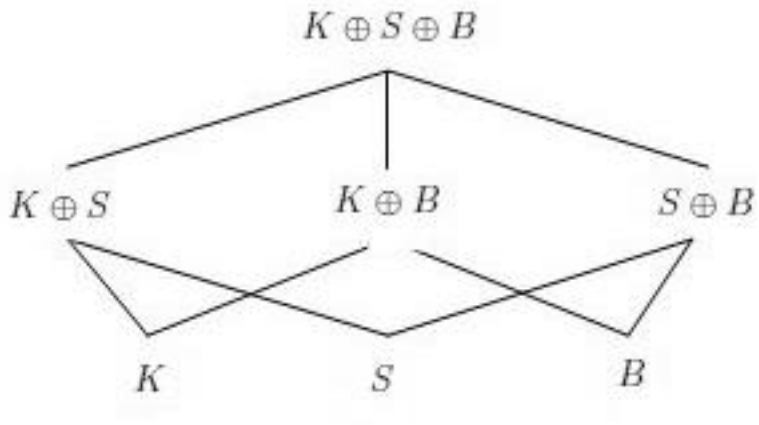

Figure 3.1: The join-semilattice of all cats in world W

The sign σ is the sum operator. σxPx represents the sum, or supremum, of all objects that are $^*P$. σ^*xPx represents the **proper sum** of Ps, that is, the supremum of all objects that are proper plural predicates of P. The difference between sum and proper sum is that the former includes (non-plural) individuals such as K or S while the latter doesn't. In worlds where there is more than one object in the extension of $^*P$, $\sigma xPx = \sigma^*xPx$: with respect to Figure 3.1, the sum of all cats is the same as the proper sum of all cats, i.e. $K \quad S \quad B$. But if we now imagine a world where there is only one cat, say Kitty, then σxPx would become K while σ^*xPx would be empty (as there is no plurality of cats in that world).

The product sign $\prod$ expresses an **individual-part** relation. The sign in combination with $\prod$ indicates a relation of **atomic part**. So we can say, for instance, that $(K \quad S) \prod (K \quad S \quad B)$: the pair *Kitty and Sylvester* is a part of the plurality *Kitty, Sylvester and Bagpuss*. Similarly, it is true that $B \quad \prod (S \quad B)$: Sylvester is an atomic part of the plurality *Sylvester and Bagpuss*.

When talking of 'stuff' rather than 'things', the sum operator σ becomes μ, the **material fusion** operator, and the individual-part operator $\prod$ becomes , the **material part** operator. So if P is the predicate *water'*, then μxPx denotes the fusion of all water in world W.

Note that in the following, as mentioned already in Section 2.1.4 and against Link's own arguments, we refer to pluralities as 'sets'. This presupposes that we treat mass terms

Bare Form	A	The
Dogs are mammals.	A dog is a mammal.	The dog is a mammal.
Ducks lay eggs.	A duck lays eggs.	The duck lays eggs.
Frenchmen eat horsemeat.	A Frenchman eats horsemeat.	?The Frenchman eats horsemeat.
Dodos are extinct.	??A dodo is extinct.	The dodo is extinct.
Mosquitoes carry malaria.	?A mosquito carries malaria.	?The mosquito carries malaria.
Birds lay eggs.	A bird lays eggs.	?The bird lays eggs.
Cars have radios.	?A car has a radio.	?The car has a radio.
Typhoons arise in this part of the Pacific.	?A typhoon arises...	?The typhoon arises...
Water is necessary for life.	??A water is necessary for life.	??The water is necessary for life.
?Chairs (in Mary's lounge) are old.	?A chair (in Mary's lounge) is old.	?The chair (in Mary's lounge) is old.

Table 3.1: The three generic constructs

as consisting of non-overlapping parts – we have already argued in favour of this view in Section 2.2.3.

3.2 Bare forms, true kinds and wide-scope stereotypes

We have already commented in the last chapter that the three constructs typical of genericity are not always substitutable (see Section 2.2.2). Table 3.1 gives some examples of possible overlaps.

For now, we will simply remark that whenever a generic can be expressed by a definite or indefinite singular, the bare plural is also available. Some generalisations, however, are not expressible through any of the accepted generic constructions (see the chair example). In the next sections, we attempt to give an interpretation of each construction which is coherent with the uses observed in Table 3.1.

3.2.1 Bare plurals and underspecified quantification

In this section, we will come back to the most ambiguous form that we have observed so far: the bare plural. We have commented in the previous chapter that it was possible to quantify bare plural sentences, albeit not always with the same quantifier:

126. Dogs are in my garden = **Some** dogs are in my garden.

127. Frenchmen eat horsemeat = **Some** Frenchmen eat horsemeat.

128. Cars have four wheels = **Most** cars have four wheels.

129. Typhoons arise in this part of the Pacific = **Some** typhoons arise in this part of the Pacific OR **Most/All** typhoons arise in this part of the Pacific.

In the current chapter, we will propose to view the ambiguity phenomena in a very slightly different light: instead of talking of ambiguous quantification, we will start talking of **underspecified quantification**, or **underquantification**. By this, we mean that the bare plural, rather than exhibiting a silent, *GEN* quantifier, simply features a placeholder in the logical form which must be filled with the appropriate quantifier. This account caters for the facts that many so-called generics can so easily be quantified via traditional quantifiers, that *GEN* is silent in all known languages, and it explains also why it is the bare form which has the highest productivity, and can refer to a range of quantified sets, from existentials to universals:

130. Dogs are in my garden. (some)

131. Dogs have four legs. (most)

132. Dogs are mammals. (all)

133. Dogs bear live young. (some?)

If the idea of describing an existential as produced by the same phenomenon as generics is uncomfortable, we refer again to Carlson (1977) where the central thesis is that bare plurals uniformly refer to kinds.

Using the underquantification hypothesis, we can paraphrase any bare plural of the form 'X does Y' as 'there is a set of things X, *a certain number of which* do Y' (note the partitive construction). This observation will become useful when we attempt to explain more difficult quantification, as in 133.

If bare plurals exhibit a quantifier slot, we must define which quantifiers can actually fill that slot. It is clear, for instance, that when we say *Cats have four legs*, we do not mean to say that *Less than 36 cats have four legs*. The quantifier *less than 36* is blocked as a slot filler for the bare plural.

There is no exhaustive list of quantifiers in English. So instead of considering each possibility, we will come back to the semantic classification of Leslie (2008) which, as we saw in Section 2.2.4, can be mapped, to a certain extent, to natural language quantifiers. To summarise, we had seen that Leslie classifies generics in three semantic classes: characteristic, majority and striking statements. We had remarked that striking statements can easily be paraphrased with the *some* quantifier, and that majority statements were basically expressions of the *most* determiner. Characteristic statements posed more of a problem, as we refused to simply formalise them with a subject/predicate structure. Let us come back for a moment to the definition of a characteristic.

While Leslie appeals to the concept of a characteristic dimension in her 2008 paper, she and her colleagues also refer to a slightly different notion in Leslie et al (2009). There, her

idea of a characteristic is compared to the notion of 'principled connection' as originally defined by Prasada and Dillingham (2006). The idea of a principled connection is that, given a subject S and a predicate P, S does P *in virtue of* being an S, or being an S *explains* why S does P. So a lion has a mane in virtue of being a lion, but it is not clear that a mosquito carries malaria in virtue of being a mosquito (Leslie et al, 2009).[1] The definition doesn't help much when it comes to quantification matters. However, we can say that its implicit vagueness is not conducive to cardinality, and *Lions have manes* will never be interpreted as *386 lions have manes*. The idea of characteristic itself doesn't imply either any sentiment (as in the striking statements) and quantifiers such as *enough* or *too many* are also out of consideration. So are negative quantifiers, which would deny a connection between subject and predicate rather than enforcing it: *(Few) lions have manes*. We therefore propose that characteristic statements are best mapped to the three basic quantifiers that we introduced in the previous chapter, i.e. *some*, *most* and *all*:

134. (Some) ducks lay eggs.

135. (Most) ducks have feathers.

136. (All) ducks are birds.

We additionally suggest that those three quantifiers are sufficient to give at least a weakly specified account of all bare plural cases: the existential reading obviously maps onto *some*.

We acknowledge that there may be some reticence in using the existential quantifier in sentences such as 134, where we are expressing more than an accidental property being associated with some accidental instances of a kind. Therefore, we spend the next two subsections considering what this partition means in terms of set quantification.

Homogeneous vs non-homogeneous predication

Although the underspecification hypothesis works well in many cases, some sentences which we have so far existentially quantified seem to call for additional semantics:

137. Turtles lay eggs. (Most female adult fertile turtles?)

138. Lions have manes. (Most adult male lions?)

[1] Prasada and Dillingham's definition of principled connection is not always helpful in making a semantic distinction between characteristic and majority statements. For instance, they take the sentence *Dogs have four legs* as expressing a principled connection. It is however dubious to see the property of having four legs as particularly characteristic of dogs. It is for instance easy to imagine that if, from tomorrow, all dogs were born with three legs as a consequence of some nuclear catastrophe, they would still be recognisable and definable as dogs.

139. Elephants have tusks. (Most African elephants and Asian male elephants?)

Sentences 137-139 belong to Leslie's characteristic statements. But as we will see next, the characterisation applies to the kind in general rather than to its instances.

When we say that *Turtles lay eggs*, we seem to make no reference to individual turtles but we simply state a fact about the mode of reproduction of the species. It takes enormous amounts of world knowledge to precisely resolve the referent of the noun phrase in those cases and this makes it impossible to believe that the grammatical construct itself could say anything about individual instances of the kind it refers to. It would therefore be tempting to just apply the term 'kind' to those noun phrases, and give the statements a subject/predicate formalisation. We will however refrain from doing so, partly for the reasons expressed in Section 2.2.3 (some individual turtles, after all, *do* lay eggs), and partly because we will reserve the term for the definite singular construct, which we take as being truly kind-referring (see Section 3.2.2). Instead, we will introduce a new notion, that of **homogeneous** predication.

There is a fundamental difference between statements such as *Turtles lay eggs* and the following:

140. Barns are red. (Khemlani et al, 2007)

141. Cats have four legs.

Although it is possible for any barn to be red, or for any cat (at birth) to have four legs, it is not possible for any given turtle to lay eggs: it must be female. We could also express this in a slightly different way, using probabilities. The probability that a given barn is red is (roughly) independent from the other features of that barn (the identity of the owner might come into play). So we can write, for instance, $P(red\ large) \quad P(red)$: the probability that a barn is red given that it is large is simply the probability for that barn to be red and is therefore the same for all barns, regardless of size. But the probability that a given turtle lays eggs is dependent on, say, its sex and its age. So $P(layEggs\ male) \ = P(layEggs\ female) \ = P(layEggs)$. We will say that the cases where independence holds display **homogeneous predication**, while the others are cases of **heterogeneous predication**.

As a side comment, we refer again to the experiments by Khemlani et al (2007) to which we briefly alluded in Section 2.2.6, where human subjects tended to wrongly accept statements such as *All ducks lay eggs*. The authors commented that those mistakes happened mostly with characteristic statements (in Leslie's terminology). Looking at the sentences that were provided during the experiments, we can ascertain that all so-called characteristic statements were actually heterogeneous. We thus suggest that when quantifying over heterogeneous predicate statements, humans might automatically perform feature

selection, as appropriate, before quantifying the noun phrase – leading to the observed mistakes.

At any rate, the problematic generics which we have identified, i.e. the ones that cannot be quantified by any other determiner than *some*, are all of the heterogeneous type, and we will show in the next subsection that it is necessary to distance ourselves from the homogeneity type of the statement when dealing purely with quantification.

Quantification, not qualification

We have shown so far that majority and striking statements could easily be mapped to a single quantifier. We have also remarked that existential quantification seems to be missing some of the semantics of the sentence in cases of characteristic heterogeneous predication. In what follows, we will argue that *some* remains, despite the doubts expressed in the last sections, the correct quantifier for such statements.

What proportion of turtles actually lays eggs? Assuming an equal split between males and females, and a higher number of (non-fertile) youngs than adults, probably around 20%. The task that we set out to achieve was described in Section 2.1.1 as the allocation of an adequate, fully specified quantifier to ambiguously quantified noun phrases. We also defined the notion of a fully specified quantifier as a quantifier for which one unique set relation exists. Given the set of all turtles T and the set of all things that lay eggs E, and the constraint that their intersection represents roughly 20% of T, it seems that the set relation $0 < \quad T \quad E \quad < \quad T \quad E \quad$ is the most appropriate for representing the cardinality of the intersection $\quad T \quad E \quad$. Using the paraphrase that we introduce earlier, we can say that we are considering the set of all possible turtles as Nbar referent and that *some of those* (the NP referent) lay eggs. This interpretation can be adopted for all similar characteristic heterogeneous statements, such as *Lions have manes* or *Peacocks have colourful tails*.

Thus, we feel entitled to argue that as far as quantification goes, an existential interpretation of those statements is accurate. If, however, we wanted to further qualify the referent of the noun phrase, that is, to apply a consistent set of attributes to those instances which are picked out by the quantifier, we would need to do more semantic work. The scope of this work being quantification rather than qualification, we do not investigate the matter any further.

We will finish this section with a brief comment on the choice of the bare plural as the preferred construct for difficult quantification. We refer once again to the broad paraphrase that we gave of the underspecified quantifier: there are Xs, *a certain number of which* are involved in Y. We note that the paraphrase is sufficiently vague to include cases that are truly ambiguous like in 142 below, cases where the speaker herself does not actually know the correct cardinality for X (we may not know whether some or most

mosquitoes carry malaria, but we can still confidently utter the sentence *Mosquitoes carry malaria*), and indeed characteristic heterogeneous statements (*Ducks lay eggs*).

142. Typhoons arise in this part of the Pacific = **Some** typhoons arise in this part of the Pacific OR **Most/All** typhoons arise in this part of the Pacific.

Underspecified quantification proves to be a useful trick for language efficiency, and a bare form seems to be appropriate to express the needed vagueness.

We have seen earlier in this chapter that bare plurals can paraphrase certain definite and indefinite singular constructs. We will consider those in the next two sections, and see whether our claim — all generics can be quantified — holds for the paraphrases.

3.2.2 True and derived kinds

We have already seen in Chapter 2 that some D-generics do not allow for direct inference at the instance level:

143. The dodo is extinct.

Quantification of such phrases seems straightforward: it can be done at the concept level in the same way that one quantifies proper nouns, i.e. the quantity inferred is one and the NP referent is the kind, the abstraction, *Dodo*. We will call those D-generics **true kinds**.

However, a subject/predicate formalisation, as we will now see, is not sufficient to give a full account of true kinds. The problem is that sentences such as 143 can actually be paraphrased as bare plurals (*Dodos are extinct*) and that we have just argued for a quantified interpretation of those. The definite singular construct, moreover, crops up in other contexts: namely, in paraphrases for the bare plurals that Leslie would qualify as 'characteristic'.

144. The duck lays eggs.

145. The lion has a mane.

146. The elephant has tusks.

Leslie et al (2009) show that generic definite singulars are strongly linked to characteristic statements and that human subjects consider majority and striking statements (*Cars have radios*, *Mosquitoes carry malaria*) far less natural in the definite singular form. This is no surprise, as we have already remarked in Section 3.2.1 that those statements express

a significant property of a kind. The noun phrases involved, though, cannot be defined as true kinds, as they allow some inference at the instance level. Saying that *The duck lays eggs* is saying that some individual ducks indeed lay eggs. So, although it could be correct to quantify such statements at the kind level, in the way we suggested for true kinds, it is not sufficient. As we saw in Section 2.2.3, telling someone that *The unicorn is a mammal* allows them to infer that individual unicorns are mammals with no more problems than if we had used a bare plural construct. Quantification is therefore at play in such singular definites and is the same as for the equivalent bare plural.

We suggest that characteristic statements slip into definite singulars because of their particular semantics and that they should be quantified both at the concept level, like true kinds, and at the instance level, like bare plurals. Because of their relation to the concept level, we will use the term **derived kinds** to qualify them.

Giving two separate formalisations to derived kinds could be seen as just an intermediary step on the way to finding a unified theory for their semantics. However, some accounts of generics actually argue for real ambiguity (see Gerstner-Link and Krifka, 1993). Cohen (2001) similarly claims that some generics are ambiguous between an inductive reading and a rules and regulations interpretation. He reworks a scenario originally proposed by Carlson (1995) where a store manager has just raised the price of bananas from \$.49/lb to \$1/lb. If we imagine, Cohen argues, that the cashiers in the store haven't realised the price increase and keep selling the bananas for \$.49, then both the following sentences are true, one under the inductive reading and the other one under the rule reading:

147. Bananas sell for \$.49/lb.

148. Bananas sell for \$1/lb.

It is easy to transfer this argument to the interpretation of derived kinds, by saying that sentence 145 both implies a biological rule under which the lion species has a mane and a statistical observation where instances of lions are seen to have a mane.

The ambiguity account, however, does not do much for true kinds and we are still at a loss when trying to explain bare plurals such as *Dodos are extinct*. For now, we will just remark that the use of a plural form indicates access to the instance-level, and we will show later in Section 3.4.2 that a formalisation is possible under the quantified reading.

3.2.3 Wide-scope stereotypes

Some of the comments that we made regarding definite singular generics can be repeated when considering indefinite singular generics. Not only are they available for characteristic statements:

149. A turtle lays eggs.

150. A lion has a mane.

151. An elephant has tusks.

They were also considered as less natural than bare plurals in majority and striking statements in the experiments by Leslie et al (2009).

One obvious difference, though, between the definite and indefinite form is that the latter is not appropriate for expressing true kinds:

152. ??A dodo is extinct.

Further, the semantics of the indefinite form is slightly different. Burton-Roberts (1977) observes what he calls 'moral necessity' in generics such as the following:

153. A gentleman opens doors for ladies.

The indefinite singular is also only conducive to one scope in statements where the corresponding bare plural would produce two different scopes (Cohen, 2001):

154. A stork has a favourite nesting area.

155. Storks have a favourite nesting area.

While the bare plural can either mean 'Storks all have a particular favourite nesting area, (namely Africa)' or 'Each stork has a favourite nesting area which is particular to that stork', the indefinite singular only produces the second scope. Because of this, and because of Burton-Roberts' remarks on the semantics of some indefinite singulars, Cohen argues that an inductivist reading of indefinite singular generics is not possible: they are always rules. Those rules, however, have a more complex formalisation than the simple subject/predicate structure suggested by Carlson (1977) and allow inferences to be made at instance level. Cohen proposes a simple formula of the type $\phi(x) \quad \psi(x)$ to account for those rules, but he declines any further comment on what the conditional actually means in such a formalisation.

As a further issue, we note that the 'moral necessity' statements of Burton-Roberts are not clearly characteristic in Leslie's sense. They seem to mostly belong to socially-constructed stereotypes of the type *A girl likes pink* or *A boy doesn't cry.* It is difficult to apply Prasada and Dillingham's constraint of 'principled connection' to those examples either as they are, precisely, not principled and can change over time. Their semantic classification is therefore unclear.

We will not attempt here to suggest a formalisation that accounts for the normative semantics of indefinite singulars such as 153. This is beyond the scope of a work on quantification. We will only argue that, as in the definite singular generic case, the hearer of an indefinite singular generic statement about unicorns should be able to make inferences about individual unicorns. This is also valid for normative statements, and we would have little trouble in quantifying 153 as *Most gentlemen open doors for ladies.* We therefore suggest processing indefinite cases as definite cases: we will use both a subject/predicate structure as an approximation for the stereotype semantics and normal quantification to express the equivalent bare plural semantics — with the caveat that the subject noun phrase should always take scope over other potential topics in the sentence. To account for the scoping and semantic particularities of those constructs, we will talk of **wide-scope stereotypes**.

3.2.4 Are bare singulars bare (or even singular)?

Back in Section 2.2.2, we included bare singulars and some bare plurals in the same class of constructs, because of the observation that they could not be paraphrased with either a definite or indefinite singular. We have now also seen, in Section 3.2.2, that bare plurals of the bare generic form tend to belong to either majority or striking statements (they correspond to those statements which are not natural in the definite/indefinite singular form in Leslie's experiments, 2009). It is however difficult to argue that bare singulars never enter characteristic statements. After all, the sentence *Water is wet* seems rather characteristic. Moreover, when mass terms can take a classifier, i.e. when there is a linguistic way to access their instances, the indefinite singular becomes natural:

156. Furniture has a practical purpose.

157. A piece of furniture has a practical purpose.

This indicates that bare singulars may not, after all, belong to bare generics but rather to the class of I-generics, expressible via the bare form and the indefinite singular. We suggest, as many authors before us (e.g. Cartwright, 1975; Link, 1983; Chierchia, 1998), that bare singulars are basically related to plurals via the concept of **non-overlapping** parts and that the use of the singular is only justified by the mass semantics of those terms. This assumption allows us to simply treat them as regular bare plurals and quantify them appropriately:

158. Water was dripping through the ceiling. (Some water).

159. Furniture has a practical purpose. (Most furniture).

160. Water consists of H_2O. (All water).

Note that Example 159 and 160 also have a stereotypical reading and should be given an ambiguous formalisation as in Section 3.2.3.

The inclusion of bare singulars in I-generics closes our overview of the typical generic constructions. We have so far shown that all constructions could take a quantified formalisation (we reserve the true kind case for the end of this chapter) and that some — those with a definite or indefinite singular paraphrase — could also be read with a simple subject/predicate semantics. Before we present an adequate, unified formalisation for the quantification effect, we pause on the last construct previously identified as ambiguous: the definite plural.

3.3 Definite plurals included

We have so far assumed that only three constructs were conducive to genericity. We would now like to point out a problem that arises when confining the genericity phenomenon to those types and argue that constructions that may not be traditionally referred to as generic belong to the same quantification phenomenon.[2]

Let us consider the interpretation of the following sentences:

161. Eight chairs are in the lounge.

162. All chairs are in the lounge.

163. Some chairs are in the lounge.

164. Chairs are in the lounge.

165. The chairs are in the lounge.

We will assume here that the predicate *to be in the lounge* is distributive. The following are logical forms for sentences 161-164. (Example 164 is assumed to be a case of unstressed *some* — the existential *sm.*)

166. $x[(8^{*}chair')(x) \quad u[u \prod x \quad inLounge'(x)]$

There is a plurality of eight chairs x, and for each atomic part u in x, u is in the lounge.

167. $u[chair'(u) \quad inLounge'(u)]$

For each chair u in the relevant set, u is in the lounge.

[2]We had better specify here that, following the argument of Section 3.2.4, we consider mass terms as plural entities, and the discussions in this section apply to them as well as to 'real' definite plurals. An example of a definite mass term can be observed in the sentence *The rice spilt out of the bag.*

168. $x[^{*}chair'(x) \quad u[u \prod x \quad inLounge'(x)]$

There is a plurality of chairs x, and for each atomic part u in x, u is in the lounge.

169. $x[^{*}chair'(x) \quad u[u \prod x \quad inLounge'(x)]$

(As for 168.)

Now, let's consider Example 165. Traditionally, such sentences have been considered to be universally quantified. The problem with the chair example is that a universal doesn't seem appropriate: it is possible to say that *The chairs are in the lounge* even though seven out of eight chairs are in the lounge. (All eight chairs are normally in the dining room and someone has just asked *Where are the chairs?*, pointing at the remaining one). Conversely, we don't feel entitled to say that *The chairs are in the lounge* if only three out of eight are there, making a mere existential reading inappropriate as well, or at least insufficient. The best possible quantified paraphrase, we argue, involves the quantifier *most*, which unfortunately is not a particularly standard reading for a definite plural.

Let's now consider the following, proposed by Dowty (1987):

170. At the end of the press conference, the reporters asked the president questions.

Dowty remarks that it is not necessary that all reporters ask questions for the sentence to be true. In fact, it is only necessary that *some of them* did. Dowty pursues:

> "The question of how many members of the group referent of a definite NP must have the distributive property is in part lexically determined and in part determined by the context, and only rarely is every member required to have these properties."

So for Dowty, the NP is referring to a 'group', i.e. the group of reporters present at the press conference. Having paraphrased the sentence as *Some reporters asked the president questions*, we can write:

171. $x[^{*}reporter'(x) \quad u[u \prod x \quad askQuestions'(u)]]$

The problem is that we have just agreed with Dowty that the NP refers to the set of reporters as a whole, and not to specific reporters. We don't want to say 'there is a small set of reporters, each of which asked a question'; we want to say 'there is a large set of reporters — all those present at the press conference — and some of them asked a question', i.e. we want to use a partitive construction. This reading seems unavailable in 171.[3]

[3]Note that the sentence in 170 is actually ambiguous. Consider it in the following context:

With regard to this problem, we follow Brogaard (2007) who gives an account of definite plurals as partitive constructions. Brogaard examines a sentence very similar to the reporters example:

172. The students asked questions.

Her argument goes as follows:

> "The sentence is true just in case there are some things$_X$ such that 'students' is true of them$_X$ and any things$_Y$ of which 'students' is true are such that they$_Y$ are some of them$_X$ and 'some of them$_X$ asked questions' is true of them$_X$."

That is, a subset Y of X is selected via the quantifier *some'* and the verbal predicate applies (distributively) to Y. If we now postulate that underquantification applies to definite plurals as it does to generics, it is easy to give an account of the various quantification phenomena observed in examples 165, 170 and 172. We can write for 170: there is a set of reporters, and a certain number of elements in that set (some reporters) asked questions —- which is our desired reading. We will show how to formalise this reading in the next section.

Note that this effect can be observed for a range of definite plurals, including possessives and demonstratives, as well as for mass terms:

173. Your employees are dedicated. (True, even if one out of fifty likes a lie-in.)

174. Those apples have turned bad. (True, even if 10% are still okay.)

175. The rice spilt out of the bag. (True, even if 3 grains are still in the bag.)

Given the range of constructs covered by what we have called 'underquantified' generics, one could ask whether the term 'genericity' is still appropriate here, or rather whether genericity and underspecified quantification should be taken as synonyms. We will however make no attempt to redefine the existing terminology. It is true that the class of 'underquantified' statements covers many constructions that have been traditionally regarded as generic, in particular majority and striking statements. It is also true that

> Reporters Smith and Jones met with the President's advisers before the press conference. They were accompanied by a cameraman and a sound engineer. At the end of the press conference, the reporters asked the President questions.

Here, the noun phrase refers to specific reporters, and it seems acceptable to apply a universal reading to it. We only concentrate here on the reading where the noun phrase's referent is the set of reporters in the press conference room.

plurality always appeals to generalisation: when we say *your books are on the table*, we use a plural because we don't want to have to say *The Genericity Book, War and Peace and Wuthering Heights are on the table.* In effect, we generalise over the three items that we don't want to list. However, we have already remarked, and will note again later in this chapter, that an account of quantification is by no means a full semantic account of noun phrases. For instance, it is not clear how definite plurals could be classified according to Leslie's model (2008). It is therefore impossible to say at this stage to which extent the semantics of generics actually overlaps with that of definite plurals. Because of this, we will keep a conservative attitude in the matter and merely point out the parallels between constructs when considering quantification itself.

3.4 Formalisation

3.4.1 Formalising collective and distributive predicates

Some verbal predicates are by nature collective in that they refer to a group as a whole and not to its instances:

176. Antelopes gather near water holes. (??Andy the antelope gathers near water holes.)

By contrast, some predicates are always distributive:

177. Three soldiers were asleep (Tom was asleep, Bill was asleep, Cornelia was asleep.)

Most verbal phrases, though, are 'mixed predicates' that accept both readings. Sometimes, the context makes one of the readings more salient and sometimes, the sentence is truly ambiguous:

178. Three soldiers stole wine from the canteen.

    (a) Tom, Bill and Cornelia went together to the canteen to steal wine.

    (b) Tom, Bill and Cornelia each stole a bottle (or several) from the canteen (perhaps at different times of the day).

The different logical forms of collective and distributive predicates are exemplified in Link (1998):

179. The Sansculottes hailed a Cordelier.

    (a) Collective reading:
    $y[Cordelier'(y) \quad hailed'(\sigma^* x \; Sansculotte'(x), y)]$

(b) Distributive reading:
$y[Cordelier'(y) \quad u[Sansculotte'(u) \quad hailed\ '(u,y)]]$

Collective predicates, as opposed to distributive predicates, can be a source of confusion when trying to directly apply quantification to an ambiguously quantified bare plural (i.e. when trying to paraphrase):

180. (All) Americans have the right to practice the religion of their choice.

181. (??some/most/all) Americans elect a new president every five years.

Quantification seems . The second sentence is however clearly not a case of true kind or stereotype either:

182. ??The/??An American elects a new president every five years.

To solve this formalisation puzzle, we refer to the reporter example (170). It was then clearly pointed out that the reading of that sentence was that out of a set of reporters, some asked questions, i.e. there was a latent partitive construct. The exact interpretation of such a sentence is then: there was a set X of reporters in the press conference room, and a subset Y of those — as selected by the quantifier *some* — asked questions: distributively, given a reporter y in Y, that reporter asked a question. Similarly, we can say that there is a set X of Americans able to vote, and a subset Y of those — which in this case is selected by the quantifier *all* and is therefore equal to X — collectively elects the president.

We can then write:

183. $X = \sigma^* x\ reporterAtPressConference'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y$
$askQuestions\ '(z)]]$

For the collective case, we just apply the verbal predicate collectively:

184. $X = \sigma^* x\ votingAmerican'(x) \quad Y[Y \prod X \quad electPresident'(Y)]$

Note that in the two examples, we have restricted the Nbar referent to the relevant set of entities. We will not investigate in this thesis how this particular reference resolution takes place.

We can then add the quantifier resolution. The set relations that we proposed in Section 2.1.2 can be expressed in terms of the Nbar set and the NP set: if $X$is the set of all Americans able to vote, Y the subset of X selected by the quantifier, and Z the set of all things that elect the president, it is clear that Y actually represents the intersection $X \quad Z$. We can thus write, using the interpretations of *some'*, *most'* and *all'* suggested in Section 2.1.2:

185. $X = \sigma^* x\, reporterAtPressConference'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad askQuestions\ '(z)] \quad 0 < \quad Y \quad < \quad X \quad Y\ ]$

186. $X = \sigma^* x\, votingAmerican'(x) \quad Y[Y \prod X \quad electPresident'(Y) \quad X \quad Y \quad = 0]$

Note that the same principle applies to mass nouns.

187. Water was dripping through the ceiling.

$X = \mu^* x\, water'(x) \quad Y[Y \ X \quad z[z \quad Y \quad dripThroughCeiling \quad '(z)] \quad 0 < \quad Y \quad < \quad X \quad Y\ ]$

188. Water consists of H_2O.

$X = \mu^* x\, water'(x) \quad Y[Y \ X \quad H \ _2O'(Y) \quad X \quad Y \quad = 0]$

We can thus write the underspecified quantifier as:

189. $X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q(Y)] \quad quant\text{-}constraint(X, Y)]$

where the quant-constraint ensures the correct cardinality of Y for various quantifiers and the predicate Q applies distributively or collectively depending on the semantics of the sentence. X denotes the Nbar referent while Y denotes the NP referent.

Examples of Formalisation

190. Dogs are in my garden:

$X = \sigma^* x\, dog'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad inMyGarden\ '(z)] \quad 0 < \quad Y \quad < \quad X \quad Y\ ]$

191. Frenchmen eat horsemeat:

$X = \sigma^* x\, Frenchman'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad eatHorseMeat\ '(z)] \quad 0 < \quad Y \quad < \quad X \quad Y\ ]$

192. The chairs are in the lounge:

$X = \sigma^* x\, chair'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad inLounge(z)] \quad X \quad Y \quad Y\ ]$
(*most* or *all*)

193. Dogs are mammals:

$X = \sigma^* x\, dog'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad mammal\ '(z)] \quad X \quad Y \quad = 0]$

194. Storks have a favourite nesting area:

$X = \sigma^* x\, stork'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad n[nestingArea\ '(n) \quad have'(z, n)]] \quad X \quad Y \quad = 0]$

$n[nestingArea'(n) \quad X = \sigma^* x\, stork'(x) \quad Y[Y \prod X \quad z[z \quad \prod Y \quad have\ '(z, n)] \quad X \quad Y \quad = 0]]$

3.4.2 Formalising kinds and stereotypes

We have shown that true kinds could be expressed via a single subject/predicate structure. We also suggested that, as well as their inductive interpretation, derived kinds could have the same formalisation as true kinds. We could simply stop here and propose the following logical form for a statement consisting of a kind ϕ and a predicate ψ:

195. $\psi(\phi)$

This, however, doesn't account for the fact that true kinds can be expressed as (presumably quantified) bare plurals (see Section 3.2.2). It also doesn't explain anaphora phenomena such as the following:

196. The dodo is extinct but Mary says she's seen one.

It is clear that if we give the true kind *the dodo* a simple subject/predicate form, we will be at a loss to say what the instance-denoting pronoun at the end of the sentence refers to. We deduce that a correct formalisation for a true kind should allow anaphoric references to instances of that kind. In order to solve this problem, we refer to the account of type-shifting given by Chierchia (1998) and Krifka (2004). In both accounts, a kind is defined as a function that returns the greatest element of the extension of the property relevant to that kind. Using Link's notation, we will say that the kind for some noun X can be represented as the plurality

197. $Kind(X) = \sigma^* x\ X'(x)$

That is, the supremum of all instances with property X'.

Making the plurality accessible allows us to refer to its instances from the anaphoric reference. We can then write:

198. $X\ =\ \sigma^* x\ dodo'(x) \quad Y[Y \prod X \quad extinct'(Y) \quad (\ Y \quad X\ =0) \quad Z[Z \prod Y \quad see'(Mary, Z) \quad (\ Z\ =1)]]$

That is, there is the set X of all dodos that were ever in existence, a set Y which happens to be X, and a set Z which selects exactly one instance out of Y (we discuss further below the validity of applying the predicate *extinct'* to a supremum). Note that it is possible to represent the anaphora as a partitive construct of the same form as that used to formalise the quantifier. The anaphora referent can be either the NP referent or the Nbar referent, depending on context. Consider the following, already quantified, example of an anaphoric reference to an Nbar:

199. Some of the glasses were broken. They were a set.

$$X = \sigma^* x\ glass'(x) \qquad Y[Y \prod X \qquad y[y \prod Y \quad broken'(y)]$$
$$(0 < Y < X Y)] \qquad Z[Z \prod X \quad set'(Z) \quad (X Z = 0)]$$

This formalisation of kinds has the effect of treating them as collectives. We argue this is acceptable insofar as the semantics of both phenomena share similarities (see Section 2.2.5). It also gives a correct interpretation of derived kinds. Saying that *The duck lays eggs* indeed implies that ducks, collectively, lay a number of eggs. This effect is perhaps more obvious in definite plural examples such as Link's (1998) *The Sansculottes hailed a Cordelier*, which sanctions both a collective and a distributive reading.

Whether predicates such as *be extinct* apply collectively to a plurality of instances rather than distributively to a single abstraction can naturally be debated. We would however argue that the collective reading is not semantically impossible and that the plausibility of the formalisation depends, in the end, on the exact semantics of the verb. We refer to the first page of our introduction for examples of philosophical problems associated with the idea of reference. In this line of argumentation, we can ask what kind of subject is selected by the predicate *be extinct*, i.e. do we need real (living) dodos to say that dodos are extinct? Or should we assume that the dodos in *Dodos are extinct* denote the set of dodos that once existed and that 'to be extinct' means 'to lack living descendants'? These are complex questions that we are in no position to answer, so we will opt for a definition of kind based on linguistic evidence only, and state that kinds are those noun phrases that can be expressed in both singular and plural form. The definition licences the supremum reading of kinds offered by Chierchia and Krifka and this, in turn, allows the interpretation of anaphora like the one in Sentence 196.

Such an interpretation, we should note, goes against some assumptions made by the literature on generics. Consider sentences such as 200:

200. The bicycle was invented in 1817.

Such a statement is prototypical of kind-referring genericity. Krifka et al (1995) claim that some verbal predicates automatically select a kind. Examples of such predicates usually include *to invent* and *to exterminate* alongside *to be extinct*. However, not only does Example 200 prevent a bare plural paraphrase (*?Bicycles were invented in 1817*) but its only relation with the instance level is that, for the bicycle to be invented, a unique bicycle had to be made in 1817. We are far from the idea of instance supremum described above.

The reason given by Krifka et al for the kind reading is that predicates such as *to be extinct* and *to invent* can never apply to singular objects. We disagree with this. As we have shown above, there is a semantics of *to be extinct* which can apply to (dead)

dodos (confirmed by the fact that the sentence is pluralisable). It seems to us that the distinction made by Krifka et al concerns objects versus concepts rather than kinds (as seen as suprema) versus individuals.

Copestake and Briscoe (1995) argue that multiple aspects of the meaning of a word should be encoded in a single lexical entry. This accounts for ambiguities such as

201. The books on the top shelf are about syntax.

where *books* refers to both the physical object and the work of a particular author.

If we assume that every word is inherently ambiguous between an object and a concept reading, we can say that *The bicycle was invented in 1817* refers to the concept of a bicycle — in fact, to a unique entity. Conversely, when we say *Bicycles were found at the bottom of the river*, we mean concrete instances of bicycles — in this case, *some bicycles*. The two sentences can be respectively formalised as:

202. $X = \sigma^* x\, bicycle'(x) \quad Y[Y \prod X \quad inventedIn1817'(Y)] \quad (\ Y \quad X \ = 0)]$ where $X \ = 1$

203. $X = \sigma^* x bicycle'(x) \quad Y[Y \prod X \quad y[y \ \prod Y \quad inRiver\ '(y)] \quad (0 < \ Y \ < \ X \ Y \ )]$

In line with our focus on quantification, we assume in this work that the referent of the noun phrase is clear (i.e. we do not posit two separate lexical forms $bicycle_{CONCEPT}$ and $bicycle_{INSTANCE}$) and remark that our proposed formalisation for quantification is applicable regardless of the exact referent.[4] (There is much more to be said on the distinction between instances and concepts but we will not discuss the matter any further in this thesis.) In this view, the denotation of the noun phrase in 200 is indeed a supremum (the supremum of all bicycle concepts in 1817) but that supremum has cardinality one. That is, the noun phrase refers to a singular, unique entity, specifiable by the quantifier *one'*. The uniqueness of the referent is demonstrated by the fact that the sentence is not pluralisable.

Note that the cardinality of a concept supremum is not necessarily one. The following sentence refers to concepts of bicycles rather than instances and indicates that there is more than one such concept:

204. Two more bicycles were invented in later years: the electric bike and the mountain bike.[5]

[4]This is in line with the theory suggested by Copestake and Briscoe for 201. This approach has the benefit to cover cases in which the ambiguity between an instance reading and a concept reading remains, as in the sentence *Ducks lay eggs* where the kind interpretation may express that ducks, collectively, lay eggs, or that the concept *Duck* has the property of laying eggs.

[5]Such uses of concepts are traditionally referred to as 'subkinds' but we will avoid the term as it clashes with our definition of kind as 'collective' supremum.

In fact, it is entirely possible to quantify concepts in the same way that we quantify instances, as in the sentence *Cats have whiskers* (paraphrasable as *All species of cats (lions, tigers, leopards, etc) have whiskers*):

205. $X = \sigma^* x\, cat'(x) \quad Y[Y \prod X \quad y[y \prod Y \quad hasWhiskers(y)] \quad X \quad Y = 0]$

3.4.3 Formalisation summary

Table 3.2 summarises the systematic relation between major noun phrase constructs, their possible quantified interpretations and the corresponding semantics in Leslie's model (2008).

NP construct	Quantified interpretation	Leslie's classification
Generic bare form	*some'*, *most'*, *all'*	Characteristic, Majority, Striking
Generic THE	*some'*, *most'*, *all'* kind (collective quantification)	Characteristic
Generic A	*some'*, *most'*, *all'* stereotype (kind)	Characteristic
Definite plurals	*some'*, *most'*, *all'*	?
Existential bare form	*some'*	–
Existential A/THE	*one'*	–

Table 3.2: Correspondence between constructs, quantification and semantics

In the next chapter, we will introduce annotation labels that match our partitioning of the quantificational space: ONE for specific, individual entities, and SOME, MOST and ALL, to match the natural language quantifiers *some'*, *most'* and *all'*. The respective formalisations are given in Table 3.3.

Label	Formalisation
ONE	$X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q(Y)] \quad X \quad Y = 0 \quad X = 1]$ OR $Q(X)$
SOME	$X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q(Y)] \quad (0 < Y < X \quad Y)]$
MOST	$X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q(Y)] \quad (X \quad Y \quad Y < X)]$
ALL	$X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q(Y)] \quad (X \quad Y = 0)]$
KIND	$X = \sigma^* x\, P'(x) \quad Y[Y \prod X \quad Q_{COLL}(Y)] \quad (X \quad Y = 0)]$

Table 3.3: Formalisations for each annotation label

In cases where the NP is a mass noun, note that, as discussed in Section 3.1, we would use the fusion operator μ and the material part operator instead of σ and $\prod$.

3.4.4 A remark on semantics

We should finish by pointing out that giving an adequate formalisation for the quantification of generics does not imply that we offer a full account of their semantics. We already mentioned in Section 2.1.1 that quantifying an ambiguous NP does not provide any information about its relation to the specificity phenomenon. We will further remark that some issues of lexical semantics remain open: in particular, we will not attempt to explain why some noun phrases can be used as kinds while others cannot (see reference to Dahl, 1975, in Section 2.2.2):

206. The cat is a mammal.

207. The coke bottle has a recognisable shape.

208. The adult whale can weigh 150 tons.

209. ??The cat is intelligent when it has blue eyes.

210. ?The cloud consists of water.

211. ??The bean is cheap on the market.

Similarly, bare plurals cannot be used as a generic (non-existential) form for all referents:

212. (Pointing at the beans on the stove) ?Beans are cooking.

Example 212 shows that definiteness is obligatory in some generalisations, in particular those where the referent is constrained in space and time. It is not clear, though, what the relation between genericity and spatio-temporal constraints actually is. The following sentence, for instance, is perfectly acceptable although it does not refer to all beans in all possible worlds:

213. (Having mentioned the special offer at the corner shop) Beans are selling for $.20/lb!

However, we have shown that all considered constructs, whether D-generics, I-generics, mixed generics or bare forms, have at least one underquantified reading. We have also shown how collective quantification can be used to formally express kinds, resulting in one unified logical form for all cases. Therefore, although we would still be unable to generate generic sentences with high accuracy, we can offer a logical interpretation of any given generic (and non-generic) text. Using the proposed framework, we will spend the rest of this thesis investigating how to automatically apply our formalisation to naturally occurring text.

Chapter 4

Quantification resolution, the human way

Most theoretical accounts of genericity, and by extension, our own approach to underspecified quantification so far, focus on well-known linguistic examples as the basis of their investigation. Although those examples are on the whole good stereotypes of the studied phenomenon, in all its various forms, and although they often highlight rare constructs that are of theoretical importance, they are not sufficient to prove the general claims that result from their study. The same examples tend to be reused across the literature, making for a restricted observation set (the phenomenon may occur in contexts that linguists have not thought about), and they are often artificially created by authors away from any 'real' linguistic context. We argue that, in order to show that a particular model holds, it is necessary to perform as large an annotation task as possible on a sufficiently heterogeneous corpus. Given enough textual data, we can make the assumption that most expressions of the phenomenon under study will be observed, and we can be sure that our data comes from natural utterances.

In this chapter, we introduce a scheme intended to help humans perform quantification resolution on naturally occurring text. The aim of building such a scheme is two-fold: not only will it allow us to make conclusions with regard to the theoretical claims made in Chapters 2 and 3, but it will provide us with a way to produce annotated corpora that can be used as training data for a machine-learning system. As we will show in Section 4.1.2, the availability of such corpora for quantification resolution is rather limited.

In what follows, we introduce some theory on linguistic annotation and propose an annotation scheme based on the idea of underspecified quantification as laid out in Chapter 3. We then report the results of an experiment involving this scheme, as used by three human annotators.

4.1 Linguistic annotation: motivation and theory

4.1.1 Linguistic motivation

The manual annotation of corpora is a common task in computational linguistics. It is usual to talk of 'annotation' generically, to cover any process that involves humans using a set of guidelines to mark some specific linguistic phenomenon in some given text. As an introduction to this chapter, we would like to argue that, when considering the aims of an annotation task and its relation to the existing linguistic literature, it becomes possible to distinguish between various types of annotation. Further, we will show that our own effort situates itself in a poorly explored relation to formal semantics and demonstrates the importance of annotation in tasks where it has been, so far, mostly absent.

The most basic type of annotation is the one where computational linguists mark large amounts of textual data with well-known and well-understood labels. The production of tree banks like the Penn Treebank (Marcus et al, 1993) makes use of undisputed linguistic categories such as parts of speech. The aim is to make the computer learn and use irrefutable bits of linguistics. (Note that, despite agreement, the representation of those categories may differ: see for example the range of available parts of speech tag sets.) This type of task mostly involves basic syntactic knowledge, but can be taken to areas of syntax and semantics where the studied phenomena have a (somewhat) clear, agreed upon definition (Kingsbury et al, 2002). We must clarify that in those cases, the choice of a formalism may already imply a certain theoretical position – leading to potential incompatibilities between formalisms. However, the categories for such annotation are themselves fixed: there is a generally agreed broad understanding of concepts such as noun phrases and coordination.

Another type of annotation concerns tasks where the linguistic categories at play are not fixed. One example is discourse annotation according to rhetorical function (Teufel et al, 2006) where humans are asked to differentiate between several discursive categories such as 'contrast' or 'weakness of approach'. In such a task, the computational linguist develops a theory where different states or values are associated with various phenomena. In order to show that the world functions according to the model presented, experimentation is required. This usually takes the form of an annotation task where several human subjects are required to mark pieces of text following guidelines inferred from the model. The intuition behind the annotation effort is that agreement between humans support the claims of the theory (Teufel, to appear). In particular, it may confirm that the phenomena in question indeed exist and that the values attributed to them are clearly defined and distinguishable. The work is mostly of a descriptive nature — it creates phenomenological definitions that encompass bits of observable language.

Our own work is similar to the latter type of annotation in that it is trying to capture a phenomenon that is still under investigation in the linguistic literature. However, it is also

different because the categories we use are fixed by language: the quantifiers *some*, *most* and *all* exist and we assume that their definition is agreed upon by speakers of English. What we are trying to investigate is whether those quantifiers should be used at all in the context of ambiguous quantification.

The type of annotation carried out in this chapter can be said to have more formal aims than the tasks usually attempted in computational linguistics. In particular, it concerns itself with some of the broad claims made by formal semantics: its model-theoretical view and the use of generalised quantifiers to formalise noun phrases.

In the introduction to this thesis, we presented our task as a work on the concept of reference. In Chapter 2, we assumed quantifiers to denote relations between sets and presented the task of quantification resolution as choosing the 'correct' set relation for a particular noun phrase in a particular sentence — implying some sort of truth value at work throughout the process: the correct set relation produces the sentence with truth value 1 while the other set relations produce a truth value of 0. What we declined to discuss, though, is the way that those reference sets were selected in natural language, i.e. we didn't make claims about what model, or models, are used by humans when they compute the truth value of a given quantified statement (see Lepore, 1983, for a critique of model theoretic semantics as being unable to do just this). The annotation task may not answer this question but it should help us ascertain to what extent humans share a model of the world — if such thing does exist.

In Chapter 3, we argued that all subject generic noun phrases could be analysed in terms of quantification. That is, an (underspecified) generalised quantifier is at work in sentences that contains such generic NPs. It is expected that if the annotation is feasible and shows good agreement between annotators, the quantification hypothesis would be confirmed. Thus, annotation may allow us to make semantic claims such as 'genericity does quantify'. Note that the categories we assume are intuitive and do not depend on a particular representation: it is possible to reuse our annotation with a different formalism as long as the theoretical assumption of quantification is agreed upon.

We are not aware of any annotation work in computational linguistics that attempts to test a particular formal theory. In that respect, the experiments presented in this chapter are of a slightly different nature than the standard research on annotation (despite the fact that, as we will show in the next section, they also aim at producing data for a language analysis system).

We will finish this section with a brief indication of what is required from the corpus used in an annotation task. Our claims about quantification should be taken as a general hypothesis, verifiable over any fragment of English language. Thus, it is above all important that the task be performed over a balanced corpus (our aim is to show that our model functions over a representative sample of natural language as a whole). The chosen corpus should therefore include several types of text (narrative, encyclopaedic, etc), ideally

produced by different writers. The number of annotated examples should also be large enough that conclusions can be drawn from the experiment. For instance, it is desirable that the class distribution obtained in the course of the annotation actually reflects the general distribution of the described phenomena.

4.1.2 NLP motivation

As far as automating the classification is concerned, a high level of human agreement in the annotation ensures that the task may be, in theory, possible. That is, given the right features and algorithm, a machine should be able to reproduce the decisions made by human subjects. If, on the other hand, humans struggle to agree, it is an indication that the task may be too complex or too poorly defined for machines to achieve. In many cases, the annotation produced will form the basis of the training and test data sets provided to the automatic classification system.

A typical way to perform automatic linguistic annotations in Natural Language Processing 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 again necessary to have a sufficiently large corpus, with a wide variety of examples, to perform such training. In the case of quantification, there is no corpus that we know of which would give us the required data. The closest contestants are the ACE corpus (2008) and the GNOME corpus (Poesio, 2000) which both focus on the phenomenon of genericity, as described in the linguistic literature. As we will see, unfortunately, neither of those corpora are suitable for use in a general quantification task.

The ACE corpus only distinguishes between 'generic' and 'specific' entities. The classification proposed by the authors of the corpus is therefore a lot broader than the one we are attempting here and there is no direct correspondence between their labels and natural language quantifiers: we have shown in Chapter 2 that genericity didn't map to a particular division of the quantificational space. Furthermore, the guidelines contradict to some extent the literature on genericity. What follows is a quotation from the ACE annotation guidelines, v6.6 24 2008.06.13:

> "A generic mention refers to a class/kind/species of objects or a typical representative of that class/kind/species and does not point to or pick out any specific individual object(s) of that class/kind/species. So if any property predicates on a generic mention, it means the entire class referred to by the mention has that property, or all/most/any members of that class have the property." (pp. 23–24)

The requirement that a generic mention be quantifiable with *all*, *most* or *any* implies that statements such as *Mosquitoes carry malaria* either refer to a class only (i.e. they are not quantified) or are not generic at all.

Further, despite the above reference to quantification, the authors seem to separate genericity and universal quantification as two antithetical phenomena:

> "Even if the author may intend to use a GEN reading, if he/she refers to all members of a set rather than the set itself, use the SPC tag.
> [All ACE annotators] are intelligent [All ACE annotators] = SPC
> [ACE annotators] are intelligent [ACE annotators] = GEN" (p. 25)

The GNOME annotation scheme is closer in essence to the literature on genericity and much more detailed than the ACE guidelines. However, the scheme distinguishes only between generic and non-generic entities, as in the ACE corpus case, and the corpus itself is limited to three genres: museum labels, pharmaceutical leaflets, and tutorial dialogues. The guidelines are therefore tailored to the domains under consideration; for instance, bare noun phrases are said to be typically generic. This restricted solution has the advantage of providing good agreement between annotators (Poesio, 2004 reports a Kappa value of 0.82 for this annotation — we come back to the definition of Kappa in Section 4.3).

The unavailability of a corpus for our task makes it necessary for us to build one. For both theoretical and practical reasons, we must ensure that the data that we produce is of high quality, i.e. that human annotators agree sufficiently when labelling the corpus. Consequently, we need a reliable measure to calculate agreement in our experiments. In what follows, we introduce in turn our corpus and the mathematical model used to evaluate our annotation.

4.2 Annotation corpus

We stressed in Section 4.1 that the choice of corpus for an annotation experiment was vital. In this work, we use a snapshot of the English version of the online encyclopaedia Wikipedia.[1] The choice was motivated by the fact that, although Wikipedia is presented as an encyclopaedia, it contains a wide variety of text ranging from typical encyclopaedic descriptions to various types of narrative texts (historical reconstructions, film 'spoilers', fiction summaries) to instructional material like rules of games. Furthermore, each article in Wikipedia is written and edited by many contributors, meaning that the requirements for speaker heterogeneity is satisfied. Finally, we would expect an encyclopaedia to contain relatively many generics, allowing us to assess how our quantificational reading fares in a real annotation task.

[1] http://www.wikipedia.org/, last accessed 16th August 2010.

Output 1 Example annotation instance

```
digraph G211 {
"TRIPLE: weed include pigra" [shape=box];
include -> weed [label="ARG1 n"];
include -> pigra [label="ARG2 n"];
invasive -> weed [label="ARG1 n"];
compound_rel -> pigra [label="ARG1 n"];
compound_rel -> mimosa [label="ARG2 n"];
"DNT INFO: lemma::include() lempos::v tense::present
(arg::ARG1 var::weed() num::pl pos::) (arg::ARG2 var::pigra() num::sg pos::)"
[shape=box];
"FILE: /anfs/bigtmp/newr1-50/page101655" [shape=box];
"ORIGINAL: Invasive weeds include Mimosa pigra, which covers 80,000 hectares of
 the Top End, including vast areas of Kakadu. " [shape=box];
}
```

In order to create our annotation corpus, we first isolated the first 100,000 pages in our snapshot and parsed them into a Robust Minimal Recursion Semantics (RMRS) representation (Copestake, 2004) using first the RASP parser (Briscoe et al, 2006) and the RASP to RMRS converter (Ritchie, 2004). We then extracted all constructions of the type Subject-Verb-Object from the obtained corpus and randomly selected 300 of those 'triples' to be annotated. Another 50 random triples were selected for the purpose of annotation training (see Section 4.5.1).

We show in Output 1 an example of an annotation instance produced by the parser pipeline.

The data provided by the system consists of the triple itself, followed by the argument structure of that triple, including the direct dependents of its constituents, the number and tense information for each constituent, the file from which the triple was extracted and finally, the original sentence in which it appeared. The information provided to annotators, as shown in Output 4 (Section 4.4.3), is directly extracted from that representation. The particular format of output allows direct visualisation of the semantic graph in Graphviz.[2] The visualisation for the triple in Output 1 is shown in Figure 4.1. (Note that the triples were not manually checked and some parsing errors may have remained.)

In the next section, we describe our parsing pipeline in more detail.

[2]http://www.graphviz.org/, last accessed 16th August 2010.

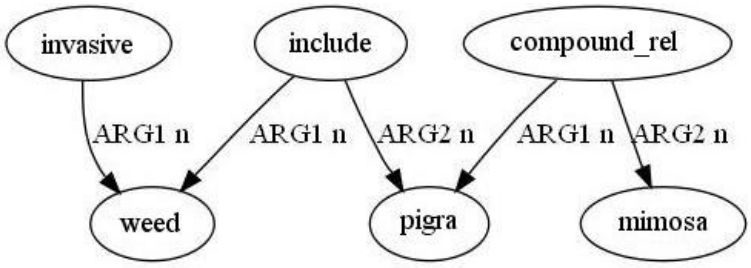

Figure 4.1: Argument graph in Graphviz

4.2.1 Parsing pipeline

The RASP parser includes tokenisation, part-of-speech tagging and lemmatisation as preprocessing steps prior to parsing. It outputs the n-best analyses of a text, displayed as syntactic trees or weighted grammatical relations. Output 2 shows an example of the trace for part-of-speech tagging together with the associated parse consisting of syntactic trees, each node headed by the relevant grammatical rule. The sentence given to the parser in this example reads: *After the first battle of Bedriacum, Vitellius became Roman emperor*. We only consider the best parse.

Output 2 Example output after RASP parsing

```
(|<w s='2' e='6'>After_ICS</w>| |<w s='8' e='10'>the_AT</w>|
 |<w s='12' e='16'>first_MD</w>| |<w s='18' e='23'>battle_NN1</w>|
 |<w s='25' e='26'>of_IO</w>| |<w s='28' e='36'>Bedriacum_NP1</w>|
 |<w s='37' e='37'>,_,</w>| |<w s='39' e='47'>Vitellius_NP1</w>|
 |<w s='49' e='54'>become+ed_VVD</w>| |<w s='56' e='60'>Roman_JJ</w>|
 |<w s='62' e='68'>emperor_NNS1</w>| |<w s='69' e='69'>._.</w>|) 1 ; (-19.575)
tree-rasp: 1
(|T/txt-sc1/-+|
 (|S/pp-np_s/+|
  (|PP/p1|
   (|P1/p_np| |<w s='2' e='6'>After_ICS</w>|
    (|NP/det_n1| |<w s='8' e='10'>the_AT</w>|
     (|N1/n-num_n1| |<w s='12' e='16'>first_MD</w>|
      (|N1/n_pp-of| |<w s='18' e='23'>battle_NN1</w>|
       (|PP/p1|
        (|P1/p_np-name| |<w s='25' e='26'>of_IO</w>|
         (|NP/n1-name|
          (|N1/n-name| |<w s='28' e='36'>Bedriacum_NP1</w>|))))))))))
  |<w s='37' e='37'>,_,</w>|
  (|S/np_vp| (|NP/n1-name| (|N1/n-name| |<w s='39' e='47'>Vitellius_NP1</w>|))
   (|V1/v_n1-tit| |<w s='49' e='54'>become+ed_VVD</w>|
    (|N1/ap_n1/-| (|AP/a1| (|A1/a| |<w s='56' e='60'>Roman_JJ</w>|))
     (|N1/n-tit| |<w s='62' e='68'>emperor_NNS1</w>|)))))
 (|End-punct3/-| |<w s='69' e='69'>._.</w>|))
```

The RMRS representation for the same sentence is shown in Output 3. RMRS is a development of Minimal Recursion Semantics (Copestake et al, 2005). One of its main features is its compatibility with both shallow and deep parsers, making it versatile enough for a wide range of applications. RMRS allows for semantic underspecification and is robust in that a structure is produced even for partial parses. In the worst case, a (highly underspecified) RMRS can be constructed from POS-tagged data alone. In this thesis, we use a compiled form of RMRS in which each sentence in the corpus corresponds to a series of minimal trees. Each tree has a root, which is one of the lemmas in the sentence, and one or more daughters, the first one of which is the index of the lemma (the other daughters being potential arguments). The elements of the trees can be co-indexed to reconstruct the whole sentence or, if the complete parse is not available, phrases in the sentence.

For the purpose of extracting triples, we convert the RMRS output into a flat representation akin to dependency relations (losing in the process some semantic information such as scopal ambiguity). This is achieved by resolving all co-indexations in the parse. For instance, the lemma *become* in Output 3 has an anchor numbered 99 which is co-indexed with the label of the second argument $ARG2$; the argument is referred to by the variable $x109$, which itself refers to the lemma *emperor*. From this information, it is possible to reconstruct the object relation of *emperor* to the verb *become*. The same process lets us identify the subject of the verb, leading to the (temporary) representation

```
lemma::become() lempos::v
(arg::ARG1 var::vitellius() num::sg pos::) (arg::ARG2 var::emperor() num::sg pos::n)
```

Note that the number and parts of speech information for the lemma and both arguments is taken directly from the parse. Arguments lacking explicit number information in the RMRS representation are assumed to be singular.

In order to obtain the representation in Output 1, we must make additional modifications to the output of the RMRS converter. First, we want to isolate the semantic information related to the triple we are annotating from the semantics for the rest of the sentence. We do this by identifying the RMRS tree that contains, as lemma, the verb of the triple, and as arguments the subject and object of that triple. We then extract direct dependents for the three words of the triple — that is, the trees that have any of the words in the triple as an argument, in any position.

Secondly, the RMRSs we are using only provide limited tense information. We use simple heuristics to recover that information from the syntactic parse. We first recover the fragment of text containing the triple, together with the associated parts of speech tags. We then make the simplifying assumption that certain patterns correspond to certain tenses and try to match the fragment to those patterns. For example:

Subject .* (has_VHZ have_VHI) .* Verb_VVN .* Object past perfect

Output 3 Example output after RMRS conversion

```
<rmrs cfrom='0' cto='74'>
<label vid='2'/>
<ep cfrom='0' cto='2'><gpred>card_rel</gpred><label vid='2'/>
    <anchor vid='3'/><var sort='u' vid='4'/></ep>
<ep cfrom='5' cto='10'><realpred lemma='after' pos='r'/><label vid='2'/>
    <anchor vid='10'/><var sort='e' vid='8'/></ep>
<ep cfrom='11' cto='14'><realpred lemma='the' pos='q'/><label vid='27'/>
    <anchor vid='26'/><var sort='x' vid='24'/></ep>
<ep cfrom='15' cto='20'><realpred lemma='first' pos='x'/><label vid='29'/>
    <anchor vid='30'/><var sort='e' vid='28'/></ep>
<ep cfrom='21' cto='27'><realpred lemma='battle' pos='n'/><label vid='25'/>
    <anchor vid='33'/><var sort='x' vid='24' num='sg'/></ep>
<ep cfrom='28' cto='30'><realpred lemma='of' pos='p'/><label vid='25'/>
    <anchor vid='36'/><var sort='e' vid='34'/></ep>
<ep cfrom='31' cto='40'><gpred>proper_q_rel</gpred><label vid='40'/>
    <anchor vid='41'/><var sort='x' vid='37' num='sg'/></ep>
<ep cfrom='31' cto='40'><gpred>named_rel</gpred><label vid='42'/>
    <anchor vid='39'/><var sort='x' vid='37'/></ep>
<ep cfrom='42' cto='51'><gpred>proper_q_rel</gpred><label vid='84'/>
    <anchor vid='85'/><var sort='x' vid='81' num='sg'/></ep>
<ep cfrom='42' cto='51'><gpred>named_rel</gpred><label vid='86'/>
    <anchor vid='83'/><var sort='x' vid='81'/></ep>
<ep cfrom='52' cto='59'><realpred lemma='become' pos='v'/><label vid='2'/>
    <anchor vid='99'/><var sort='e' vid='97' tense='present'/></ep>
<ep cfrom='60' cto='65'><realpred lemma='roman' pos='j'/><label vid='2'/>
    <anchor vid='102'/><var sort='e' vid='100'/></ep>
<ep cfrom='66' cto='73'><realpred lemma='emperor' pos='n'/><label vid='2'/>
    <anchor vid='111'/><var sort='x' vid='109'/></ep>
<rarg><rargname>ARG1</rargname><label vid='3'/><var sort='x' vid='1'/></rarg>
<rarg><rargname>CARG</rargname><label vid='3'/><constant>69</constant></rarg>
<rarg><rargname>RSTR</rargname><label vid='26'/><var sort='h' vid='76'/></rarg>
<rarg><rargname>BODY</rargname><label vid='26'/><var sort='h' vid='77'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='36'/><var sort='x' vid='24'/></rarg>
<rarg><rargname>ARG2</rargname><label vid='36'/><var sort='x' vid='37'/></rarg>
<rarg><rargname>RSTR</rargname><label vid='41'/><var sort='h' vid='43'/></rarg>
<rarg><rargname>BODY</rargname><label vid='41'/><var sort='h' vid='44'/></rarg>
<rarg><rargname>CARG</rargname><label vid='39'/><constant>bedriacum</constant></rarg>
<rarg><rargname>ARG1</rargname><label vid='99'/><var sort='x' vid='81'/></rarg>
<rarg><rargname>RSTR</rargname><label vid='85'/><var sort='h' vid='87'/></rarg>
<rarg><rargname>BODY</rargname><label vid='85'/><var sort='h' vid='88'/></rarg>
<rarg><rargname>CARG</rargname><label vid='83'/><constant>vitellius</constant></rarg>
<rarg><rargname>ARG2</rargname><label vid='99'/><var sort='x' vid='109'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='102'/><var sort='x' vid='109'/></rarg>
<hcons hreln='qeq'><hi><var sort='h' vid='76'/></hi><lo><label vid='25'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='43'/></hi><lo><label vid='42'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='87'/></hi><lo><label vid='86'/></lo></hcons>
</rmrs></rmrs-list>
```

If the subject of the triple precedes the auxiliary *have* and the auxiliary precedes the verb of the triple in the past participle form, then the tense is past perfect. We categorise all verbs into six classes: present, simple past, past perfect, future, progressive and present participles.

4.3 Evaluation of annotation agreements

We said, at the end of Section 4.1.2, that the quality of our annotation data had to be evaluated using an adequate measure. We now introduce the theoretical tools used in this thesis to compute agreement between annotators.

In an annotation task, two aspects of agreement are important when trying to prove or refute a particular linguistic model: stability and reproducibility (Krippendorf, 1980). Reproducibility refers to the consistency with which humans apply the scheme guidelines, i.e. to the so-called **inter-annotator agreement**. Stability relates to whether the same annotator will consistently produce the same annotations at different points in time. The measure for stability is called **intra-annotator agreement**. Both measures concern the repeatability of an annotation experiment.

In this work, agreement is calculated for each pair of annotators according to the Kappa measure (Cohen, 1960):

$$\kappa = \frac{Pr(a) \quad Pr(e)}{1 \quad Pr(e)} \tag{4.1}$$

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 $Pr(e)$: we use Fleiss' Kappa (Fleiss, 1971), which allows us to compute agreement between multiple annotators.

In Fleiss' Kappa, $Pr(e)$ is calculated as follows. We take A to be the number of annotators involved in the task, I the number of instances to be annotated and C the number of categories used for the classification. n_{ij} is the number of annotators who annotated instance number i to the jth category.

For each category $cat_{1...C}$ we calculate what proportion p_j of all annotations this category covers:

$$p_j = \frac{1}{AI} \sum_{i=1}^{I} n_{ij} \tag{4.2}$$

For each instance $inst_{1...I}$, we then compute p_i, the proportion of agreement amongst annotators for that instance:

Kappa Value	Kappa Interpretation
< 0	No agreement
0 - 0.2	Very low agreement
0.2 - 0.4	Low agreement
0.4 - 0.6	Moderate Agreement
0.6 - 0.8	Full Agreement
0.8 - 1.0	Perfect Agreement

Table 4.1: An interpretation of the values of Kappa. Landis and Koch (1977).

$$p_i = \frac{1}{A(A \quad 1)} \sum_{j=1}^{C} n_{ij}(n_{ij} \quad 1) \tag{4.3}$$

$Pr(a)$ is the mean of all p_is:

$$Pr(a) = \frac{1}{I} \sum_{i=1}^{I} p_i \tag{4.4}$$

$Pr(e)$ is the sum of all squared p_js:

$$Pr(e) = \sum_{j=1}^{C} p_j^2 \tag{4.5}$$

The values produced by Kappa range from negative figures to 1. In order to relate the different values to a judgement about the quality of the annotation, it is useful to refer to some agreed-upon scale. Landis and Koch (1977) provide an interpretation of Kappa which is commonly used and that we summarise in Table 4.1.

In the rest of this chapter, we introduce an annotation scheme tailored to the task of quantification, with the dual aim of providing the required training and test data to a machine learning system and to put to the test the theoretical claims introduced in Chapter 3.

4.4 An annotation scheme for quantification resolution

4.4.1 Theoretical claims: a reminder

In Chapter 3, we made the following claims:

All bare plurals and definite plurals (including definite mass terms) can be quantified across the space of *some'*, *most'* and *all'*.

True kinds can be formalised as a collective over *all* instances of that kind or, in the case of conceptual kinds, as a simple, unique entity.

Derived kinds (definite singulars) and stereotypes (indefinite singulars) accept both a bare plural and a true kind formalisation.

Bare singulars can be read and formalised as bare plurals or, in cases where they have a stereotypical reading, as stereotypes.

Collective and distributive readings lead to different formalisations of the verbal predicate in the sentence within the underspecified quantifier scope.

Making those claims has several consequences. First, we can say that any noun phrase lacking an explicit quantifier can be quantified as *some*, *most*, *all* or *one*, the latter for cases that refer to a unique entity as in *The cat is sleeping by the fire*. Secondly, we note that the exact grammatical construct of the noun phrase (i.e., whether it is singular or plural, bare or not, definite or indefinite), as well as its possible paraphrases, have an influence on its potential readings: for instance, a bare plural that can be paraphrased as indefinite singular will be a true or derived kind; similarly, an indefinite singular that cannot be paraphrased as a bare plural is just a unique, specific entity. Thirdly, if we set aside the case of collectively formalised true kinds, we observe that whether the noun phrase is a kind (stereotype) or not is in theory independent from its actual quantification (any combination of kind and quantification values is possible) and similarly, whether the verbal predicate is collective or distributive is uncorrelated with the kind status and the quantification of the NP. This has implications for the complexity of the annotation scheme.[3]

4.4.2 Scheme structure

A complete annotation scheme for the quantification resolution task can be found in Appendix A of this thesis. The scheme consists of five parts. The first two present the annotation material and the task itself. Some key definitions are given. The following part describes the various quantification classes to be used in the course of the annotation. Participants are then given detailed instructions for the labelling of various grammatical constructs. Finally, in order to keep the demand on the annotators' cognitive load to a

[3]We will see in Section 5.3.2 that this independence assumption does not hold statistically, but we do not wish to make theoretical claims in that respect. For the sake of clarity, the experiments presented in this thesis cover the collective/distributive distinction, kind value and quantification phenomenon separately.

minimum, the last part reiterates the annotation guidelines in the form of diagrammatic decision trees.

In the next sections, we give a walk-through of the guidelines and definitions provided.

4.4.3 Material

Our annotators are first made familiar with the material provided to them. This material consists of 300 entries of the type shown in Output 4. Each entry consists of a single sentence and of a triple Subject-Verb-Object which helps the annotator identify which subject noun phrase in the sentence they are requested to label. No other context is provided. This is partly to make the task shorter (letting us annotate more instances) and partly to allow for some limited comparison between human and machine performance (by restricting the amount of information given to our annotators, we force them – to some extent – to use the limited information that would be available to an automatic quantification resolution system, e.g. syntax).

Output 4 An annotation entry, as provided to human subjects

```
********* Annotation 211 *****************************************

TRIPLE:   weed include pigra
ORIGINAL: Invasive weeds include Mimosa pigra, which covers 80,000
          hectares of the Top End, including vast areas of Kakadu.

************************************************************************
```

4.4.4 Definitions

A good annotation scheme must provide definitions of key concepts in a way that a non-linguist can comprehend. In our scheme, we introduce the annotators to the three concepts of **quantification**, **kind**, and **distributivity** (versus **collectivity**).

Quantification is described in simple terms, as the process of 'paraphrasing the noun phrase in a particular sentence using an unambiguous term expressing some quantity'. An example is given:

214. *Europeans* discovered the Tuggerah Lakes in 1796 = *Some Europeans* discovered the Tuggerah Lakes in 1796.

We only allow the three quantifiers *some*, *most* and *all*. In order to keep the number of classes to a manageable size, we introduce the additional constraint that the process of quantification must yield a single quantifier. We force the annotator to choose between the three proposed options and introduce priorities in cases of doubt: *most* has priority over *all*, *some* has priority over the other two quantifiers. This ensures we keep a conservative attitude with regard to inference (see Chapter 1).

Kinds are defined with the sole purpose of making the process transparent throughout (annotators never actually need to make a direct decision as to whether a noun phrase is a kind or not). They are presented as denoting 'the group including all entities described by the noun phrase under consideration'.

Distributivity and collectivity are introduced together as a binary distinction. Distributive statements are described as those 'where every entity referred to by the subject is individually involved in the verb's action'. Collective statements, by contrast, are those 'where the group referred to by the subject, as opposed to individuals, performs the action'. Two examples are given:

215. *The students* took an exam = each student, individually, took the exam.

216. *The residents* founded a self-help group = the residents, together, founded a self-help group (and not: each resident founded their separate self-help group).

Because statements are very often ambiguous with regard to collectivity and distributivity, we request that annotators only label a verbal predicate as collective if a distributive reading is totally impossible, or at least extremely unlikely, as in 216.

Quantification classes are introduced in a separate part of the scheme. We define the five labels ONE, SOME, MOST, ALL and QUANT (for already quantified NPs) and give examples for each one of those.

We try, as much as possible, to keep annotators away from performing complex reference resolution. When quantifying, their first task is therefore to simply attempt to paraphrase the existing sentence by appending a relevant quantifier to the noun phrase to be annotated. In some cases, however, this is impossible and no quantifier yields a correct English sentence (this often happens in collective statements). To help our annotators make decisions in those cases, we ask them to distinguish what the noun phrase might refer to when their first hear it and what it refers to at the end of the sentence, i.e., when the verbal predicate has imposed further constraints on the quantification of the NP.

4.4.5 Guidelines and decision trees

Guidelines are provided for five basic phrase types: quantified noun phrases, proper nouns, plurals, non-bare singulars and bare singulars.

Quantified noun phrases

This is the simplest case: a noun phrase that is already quantified such as *some people*, *6 million inhabitants* or *most of the workers*. The annotator simply marks the noun phrase with a QUANT label.

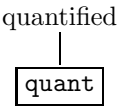

Figure 4.2: Quantified case.

Proper nouns

Proper nouns are another simple case. However, because what annotators understand as a proper noun varies, we provide a definition. We note first that proper nouns are often capitalised. It should however be clear that, while capitalised entities such as *Mary*, *Easter Island* or *Warner Bros* refer to singular, unique objects, others refer to groups or instances of those groups: *The Chicago Bulls*, *a Roman*. The latter can be straightforwardly quantified:

214. The Chicago Bulls won last week. (ALL, collective)[4]

[4]We acknowledge that treating team names as quantified entities can be an issue. In particular, it causes issues of consistency. Having annotated 214 as ALL, we would expect to do the same for the subject noun phrase in *Manchester United won last week*. However, for the annotation to make sense, we must assume that the reader automatically associates the players in the team with the team itself, which in many cases is a false generalisation (consider *Manchester United was founded in 1878*). Sag et al (2002) actually regard US team names as semi-fixed multi-word expressions introduced by a definite specifier and automatically followed by a plural. Under this view, we should annotate 214 as ONE. We argue, however, that some team names behave differently from others: for instance, it is possible to utter *The Chicago Bulls are all proud of their team* but the equivalent, *Manchester United are all proud of their team*, is less satisfactory. We argue that the relation between Manchester United and its players is metonymical while the relation between the Chicago Bulls and individual Bulls is not.

215. A Roman shows courage in battle. (MOST, stereotype, distributive)

We define proper nouns as noun phrases that 'contain capitalised words and refer to a concept which doesn't have instances'. All proper nouns are annotated as ONE.

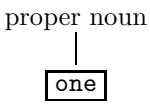

Figure 4.3: Proper noun case.

Plurals

Plurals require several decisions to be made: not only must they be appropriately quantified and the distributive/collective status of their verbal predicate indicated, but the annotators must also specify whether they are kinds or not. This last decision can simply be made by attempting to paraphrase the sentence with either a definite singular (leading to a true or derived kind) or an indefinite singular (leading to a stereotype).

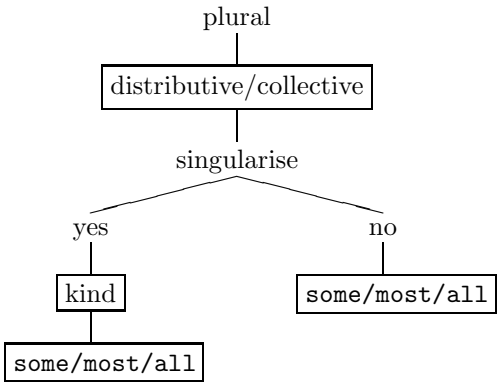

Figure 4.4: Plural case.

(Non-bare) singulars

Like plurals, singulars must be tested for a kind reading. Not surprisingly, this is done by attempting to pluralise the noun phrase. If pluralisation is possible, then the kind interpretation is confirmed and quantification is performed. If not, the singular refers to a single entity and is annotated as ONE.

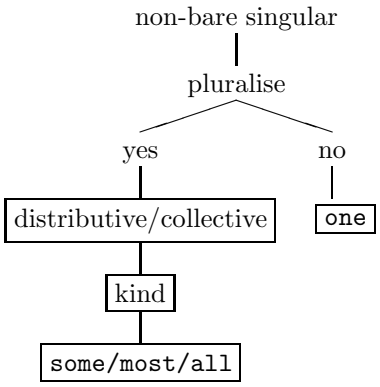

Figure 4.5: Non-bare singular case.

Bare singulars

In Chapter 3, we described bare singulars as essentially plural, under the linguistic assumption of non-overlapping atomic parts. In order to make this relation clear, we ask annotators to try and paraphrase bare singulars with an (atomic part) plural equivalent and follow, as normal, the decision tree for plurals. An example is supplied:

216. *Free software* allows users to co-operate in enhancing and refining the programs they use *Open source programs* allow users to co-operate...

When the paraphrase is impossible, the noun phrase is deemed a unique entity and annotated as ONE.

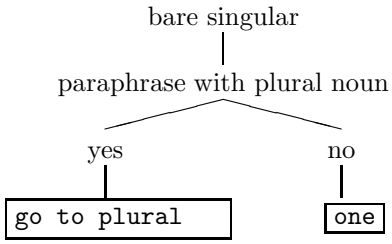

Figure 4.6: Bare singular case.

4.5 Implementation and results

4.5.1 Task implementation

Three annotators were used in our experiment. One annotator was the author of this thesis (subsequently referred to as 'Annotator 1' in all results tables); the other two annotators were graduate students (non-linguists), both fluent in English. The two graduate

students were provided with individual training sessions where they first read the annotation guidelines, had the opportunity to ask for clarifications, and subsequently annotated, with the help of the author, the 50 noun phrases in the training set. Both training sessions ran in a similar fashion and lasted about 90 minutes. After the initial familiarisation with the guidelines, the annotators were invited to go through the training set. For each annotation instance, the answer given by the annotator was immediately checked against that given by the author. When disagreement occurred, the annotator was asked to provide a justification for the label they had chosen. In cases where the annotator's interpretation of the guidelines differed from that intended by the author, clarification was given by the author: this process ensured guidelines were used as consistently as possible amongst annotators.

Subsequently, the 300 noun phrases selected for the actual annotation task were labelled by all three annotators individually. This phase involved no communication between them.

4.5.2 Kappa evaluation

As we made an independence assumption between quantification value, kind value and distributivity value (see Section 4.4.1), we can evaluate our annotator agreement separately for each type of annotation.

Intra-annotator agreement

Intra-annotator agreement was calculated over the set of annotations produced by the author of this thesis. The original annotation experiment was reproduced at three months' interval and Kappa was computed between the original set and the new set. Table 4.2 shows results over 0.8 for all three tasks, corresponding to 'perfect agreement'. This demonstrates that the stability of the scheme is high.

Class	Quantification	Kind	Distributive/Collective
Kappa	0.84	0.85	0.86

Table 4.2: Intra-annotator agreements for all three tasks

Inter-annotator agreement

Table 4.3 shows inter-annotator agreements of over 0.6 for all three tasks, which correspond to 'substantial agreement' according to the Landis and Koch classification. This result must be taken with caution, though. Although it shows good agreement overall,

Class	Quantification	Kind	Distributive/Collective
Kappa	0.72	0.67	0.65

Table 4.3: Inter-annotator agreements for all three tasks

it is important to ascertain in what measure it holds for separate classes. In an effort to report such per class agreement, we calculate Kappa values for each label by evaluating each class against all others collapsed together (as suggested by Krippendorf, 1980).

Tables 4.7 and 4.9 indicate that substantial agreement is maintained for separate classes in both the kind and distributive/collective annotation tasks (with the exception of the collective class, which scores just below the 0.6 threshold). Table 4.5, however, suggests that, if agreement is perfect for the ONE and QUANT classes, it is very much lower for the SOME, MOST and ALL classes. While it is clear that the latter three are the most complex to analyse, we can show that the lower results attached to them are partly due to issues related to Kappa as a measure of agreement. Feinstein and Cicchetti (1990), followed by Di Eugenio and Glass (2004) proved that Kappa is subject to the effect of prevalence and that different marginal distributions can lead to very different Kappa values for the same observed agreement. It can be shown, in particular, that an unbalanced, symmetrical distribution of the data produces much lower figures than balanced or unbalanced, asymmetrical distributions because the expected agreement gets inflated. The confusion matrices in Tables 4.6, 4.8 and 4.10 indicate that our data falls into the category of unbalanced, symmetrical distribution: the classes are not evenly distributed but annotators agree on the relative prevalence of each class. Moreover, in the quantification task itself, the ONE class covers roughly 50% of the data. This means that, when calculating per class agreement, we get an approximately balanced distribution for the ONE label and an unbalanced, but still symmetrical, distribution for the other labels.

Class	Kappa	Pr(a)	Pr(e)
ONE	0.814	0.911	0.521
SOME	0.445	0.893	0.808
MOST	0.438	0.931	0.877
ALL	0.509	0.867	0.728
QUANT	0.884	0.987	0.885

Table 4.4: The effect of prevalence on per class agreement, quantification task.

Class	ONE	SOME	MOST	ALL	QUANT
Kappa	0.81	0.45	0.44	0.51	0.88

Table 4.5: Per class inter-annotator agreement for the quantification annotation

Class	ONE_1	$SOME_1$	$MOST_1$	ALL_1	$QUANT_1$	
ONE_2	173	1	1	3	0	178
$SOME_2$	5	22	6	10	0	43
$MOST_2$	3	1	12	7	0	23
ALL_2	6	3	1	26	1	37
$QUANT_2$	2	0	0	0	17	19
	189	27	20	46	18	300

Class	ONE_1	$SOME_1$	$MOST_1$	ALL_1	$QUANT_1$	
ONE_3	167	2	0	5	1	175
$SOME_3$	5	12	3	6	1	27
$MOST_3$	3	2	8	3	0	16
ALL_3	13	10	9	32	0	64
$QUANT_3$	1	1	0	0	16	18
	189	27	20	46	18	300

Class	ONE_2	$SOME_2$	$MOST_2$	ALL_2	$QUANT_2$	
ONE_3	163	5	2	3	2	175
$SOME_3$	3	15	5	3	1	27
$MOST_3$	3	4	8	1	0	16
ALL_3	9	18	8	29	0	64
$QUANT_3$	0	1	0	1	16	18
	178	43	23	37	19	300

Table 4.6: Confusion matrices for the quantification annotation. The indices identify the annotators.

Class	KIND	NOTKIND	QUANT
Kappa	0.63	0.71	0.88

Table 4.7: Per class inter-annotator agreement for the kind annotation

Class	NOTKIND_1	KIND_1	QUANT_1	
NOTKIND_2	20	17	0	37
KIND_2	6	237	1	244
QUANT_2	0	2	17	19
	26	256	18	300

Class	NOTKIND_1	KIND_1	QUANT_1	
NOTKIND_3	21	11	1	33
KIND_3	5	243	1	249
QUANT_3	0	2	16	18
	26	256	18	300

Class	NOTKIND_2	KIND_2	QUANT_2	
NOTKIND_3	23	9	1	33
KIND_3	14	233	2	249
QUANT_3	0	2	16	18
	37	244	19	300

Table 4.8: Confusion matrices for the kind annotation. The indices identify the annotators.

Class	DIST	COLL	QUANT
Kappa	0.67	0.57	0.88

Table 4.9: Per class inter-annotator agreement for the distributive/collective distinction

Class	DIST_1	COLL_1	QUANT_1	
DIST_2	232	9	0	241
COLL_2	13	26	1	40
QUANT_2	2	0	17	19
	247	35	18	300

Class	DIST_1	COLL_1	QUANT_1	
DIST_3	227	10	2	239
COLL_3	19	24	0	43
QUANT_3	1	1	16	18
	247	35	18	300

Class	DIST_2	COLL_2	QUANT_2	
DIST_3	222	14	3	239
COLL_3	18	25	0	43
QUANT_3	1	1	16	18
	241	40	19	300

Table 4.10: Confusion matrices for the distributive/collective distinction. The indices identify the annotators.

This leads to the expected agreement being rather low for the ONE class and very high for the other classes. Table 4.4 reproduces the per class agreement figures obtained for the quantification task but shows, in addition, the observed and expected agreements for each label. Although the observed agreement is consistently close to, or over, 0.9, the Kappa values differ widely in conjunction with expected agreement. This produces relatively low results for SOME, MOST and ALL (the QUANT label has nearly perfect agreement and therefore doesn't suffer from prevalence). The same consequences can be drawn for the kind annotation and the distributive/collective distinction.

We also note that the quality of the annotation provided by Annotator 3 seems slightly worse than that achieved by the other two annotators. Looking at Table 4.6, we see that the confusion matrix obtained by Annotators 1 and 2 displays a strong diagonal. It is however not the case in the matrices involving Annotator 3. Further training might have been beneficial for that annotator.

With regard to the purpose of creating a gold standard for the quantification resolution task, we note that out of 300 quantification annotations, there are only 14 cases in which a majority decision cannot be found, i.e., at least two annotators agreed in 95% of cases. Thus, despite some low Kappa results, we believe that the data can adequately be used for the production of training material. (As far as such data ever can be: Reidsma and Carletta, 2008, show that systematic disagreements between annotators will produce bad machine learning, regardless of the Kappa obtained on the data.)

In Section 4.5.3, we introduce some of the difficulties encountered by our subjects, as related in post-annotation discussions. We focus on quantification only, as agreement for the distributive/collective distinction and the kind annotation are satisfactory.

4.5.3 Annotation issues

Several issues came up in the course of the annotation. We identified, in particular, three areas of difficulty: the choice of referent for the noun phrase, the need for world knowledge in certain cases and the interaction between quantification over instances and quantification over situations. We review those three areas in turn.

Reference

Although we tried to make the task as simple as possible for the annotators by asking them to paraphrase the sentences that they were reading, they were not free from having to work out the referent of the noun phrase (consciously or unconsciously) and we have evidence that they did not always pick the same referent, leading to disagreements at the quantification stage. Consider, for instance, the following sentence:

217. Subsequent annexations by Florence in the area have further diminished the likelihood of incorporation, and no serious attempts have been made since.

In the course of post-annotation discussions, it became clear that not all annotators had chosen the same referent when quantifying the subject noun phrase in the first clause. One annotator had chosen as referent *subsequent annexations*, leading to the reading *Some subsequent annexations, conducted by Florence in the area, have further diminished the likelihood of incorporation...* The other two annotators had kept the whole noun phrase as referent, leading to the reading *All the subsequent annexations conducted by Florence in the area have further diminished the likelihood of incorporation...*

World knowledge

Being given only one sentence as context for the noun phrase to quantify, annotators sometimes lacked the world knowledge necessary to make an informed decision. The following sentence illustrates the problem:

218. The undergraduate schools maintain a nonrestrictive Early Action admissions programme.

It came to light that all three annotators had a different interpretation of what the mentioned Early Action programme might refer to, and of the duties of the undergraduate schools with regard to it. This led to three different quantifications: SOME, MOST and ALL.

Interaction with time

The existence of interactions between noun phrase quantification and what we will call temporal quantification is not surprising: we refer to the literature on genericity and in particular to Krifka et al (1995) who talk of characteristic predication, or habituality, as a phenomenon encompassed by genericity (see Section 2.2.2). We do not intend to argue for a unified theory of quantification, as temporal quantification involves complexities which are beyond the scope of this work. However, the interaction between temporality and noun phrase quantification might explain further disagreements in the annotation task. In what follows, we describe two such cases in some detail.

The first case of disagreement involves a sentence with a temporal adverb:

219. Scottish fiddlers emulating 18th-century playing styles sometimes use a replica of the type of bow used in that period.

Two annotators labelled the subject of that sentence as MOST, while the third one preferred SOME. In order to understand the issue, consider the following, related, statement:

220. Mosquitoes sometimes carry malaria.

This sentence has the possible readings: *Some mosquitoes carry malaria* or *Mosquitoes, from time to time in their lives, carry malaria.* The first reading is clearly the preferred one.

The structure of Sentence 219 is identical to that of Sentence 220 and it should therefore be taken as similarly ambiguous: it either means that some of the Scottish fiddlers emulating 18th-century playing styles use a replica of the bow used in that period, or that a Scottish fiddler who emulates 18th-century playing styles, from time to time, uses a replica of such a bow. The two readings correspond to the labels given to that sentence by the annotators. We could however argue that when the temporal adverb is used at quantification stage — i.e., when this is the preferred reading for the sentence – the noun phrase should actually be annotated as already quantified. The ambiguity in Sentence 219 thus involves the labels QUANT and MOST rather than SOME and MOST.[5]

The second case of disagreement involves both a temporal adverb and a conditional construction. But it can be analysed in an identical way:

221. This often happens when a player is holding three cards of one suit and draws from the deck, picking up a fourth of that suit accidentally.

When asked to annotate the noun phrase *a player*, two out of three annotators judged the sentence pluralisable without loss of meaning. The third annotator argued that the statement refers to an illustrative (imaginary) situation and that pluralisation would imply, falsely, that several players in that particular situation are involved in the verbal predicate. The two annotators who pluralised the sentence admitted problems when quantifying. One chose the label MOST, the other one chose ALL.

To make the following argument clearer, we will simplify the sentence:

222. When a player holds three cards of one suit, she often wins tricks.

The new sentence can be paraphrased as

[5]Note that when the temporal adverb is made to quantify the noun phrase, the verbal phrase still needs to be temporally quantified.

> Some mosquitoes (sometimes) carry malaria.
>
> Some Scottish fiddlers emulating 18th-century playing styles (always?) use a replica of the bow used in that period.

223. A player holding three cards of one suit often wins tricks.

The paraphrase very much resembles Sentence 219. It is similarly ambiguous and can be taken as meaning *Most player holding three cards of one suit win tricks* or *A player holding three cards of one suit will often win tricks/will win many tricks in one given game*. The first reading is the 'already quantified' reading while the second one, after pluralisation, is readable as involving quantification via ALL (assuming that the rules of the game favour a player who holds many cards of one suit). Again, the disagreement between the two plural quantifiers becomes clear. (We would argue that the ONE quantification is altogether incorrect as it does not reflect the generalisation expressed by the sentence as a whole).

4.6 Advantages over genericity annotation

We finish this chapter with a short, qualitative comparison between our quantification annotation scheme and our previous efforts in annotating genericity. We claim that the solution presented for quantification is more viable for large datasets in terms of ease of annotation and reliability of performance.

In Herbelot and Copestake (2008), we developed two sets of guidelines to distinguish between generic and non-generic entities. The first scheme aimed to classify noun phrases under four labels, SPEC, GEN, GROUP and NON-WE, corresponding respectively to specific entities, generics, generics in need of further reference resolution and entities that were not 'well established' in a sense close to that proposed by Krifka et al (1995). Examples, with further explanations, are given below:

224. (a) *The cat* is sleeping by the fire. SPEC
= A single, individual cat is sleeping by the fire.

(b) *The whale* feeds on plankton. GEN
= Typically, a whale feeds on plankton.

(c) The whale feeds on plankton. *The animal* lives on extensive fat reserves. GROUP
= The whale (as denoted by the anaphora *the animal*) lives on extensive fat reserves — typically, a whale lives on extensive fat reserves.

(d) *Elaborate schemes of still greater complexity* are sometimes used in the field of cryptography. NON-WE
(The subject noun phrase has no clear ontological status: it might be difficult to agree on the definition of an 'elaborate scheme of still greater complexity'.)

The scheme relied on the ability of annotators to conceptualise the referent of noun phrases. They were asked, for instance, whether the entity was distinguishable from similar entities in the actual world (in an effort to ascertain their individual, as opposed to generic, status), or whether they could think of hyponyms/instances for the concept. The Kappa values obtained for the scheme over 100 noun phrases, as annotated by six graduate students, lay in the higher range of low agreement and the lower range of moderate agreement. The error analysis performed on the results showed that humans had difficulties in explicitly resolving the denotation of the noun phrase as well as the semantics of its verbal predicate. In order to remedy to this problem, a second scheme was produced, which attempted to break down the decision process in smaller, more manageable steps.

The second scheme went through several development iterations. The final version contained 14 steps and catered for specific cases such as existentials, proper nouns and copula constructions. The main additions are explained below:

A step was dedicated to referent resolution. The annotator was 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. (Context was provided).

Unique entities were dealt with in two steps, 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, had showed that the concept of 'cryptography' had been marked as specific by some annotators in contexts where genericity would have prevailed, because 'there are not several cryptographies'.

A label for non-specifics was also added: 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. The new requirement was that entities that refer to a particular object be classified as either specific (identifiable) or nonspecific (non-identifiable). We used Jørgensen's definition of specificity as given in Section 2.2.5.

The differentiation between groups and generics was made simpler by comparing the textual entity P with its referent resolution $P2$ (as performed at the beginning of the annotation process): when $P = P2$, the entity was deemed generic, otherwise a group.

Finally, there was a new, explicit ambiguity label which could be applied to bare plurals. The annotator was requested to reconsider non-specific entities in bare plurals and asked whether there was a reading of the sentence where the entity might refer to a class of objects.

The new scheme was fairly complex, and although it produced a better Kappa in a restricted experiment over 50 noun phrases, it was not clear that the guidelines could be efficiently ported to large-scale annotation tasks. Further, problems related to referent resolution remained. As shown in Section 4.5.3, our quantification annotation scheme has potential for further improvement. However, there is a clear qualitative gain over the genericity annotation task. First, the annotation process is much faster (the author, comparing her performance at the end of the guidelines development process in both tasks, estimates the speed increase to be threefold). Second, the scheme uses a set of clear lexical items as labels and avoid complex linguistic concepts, the definition of which is liable to change as theories evolve. Third, it doesn't necessitate reference resolution in all cases and some of the issues caused by reference can be solved without the need to define the concept: instructing the annotators to always quantify the whole noun phrase, with dependents, would help avoid issues such as the one encountered in Sentence 217.

We should note that the resources available for this project did not allow us to fully investigate the process gone through by the annotators when performing quantification resolution. We are assuming that our guidelines were used consistently but, as pointed out by Schütze (2005), this does not ensure that we fully understand the interpretation that annotators gave to those guidelines. For instance, we do not know at which point annotators used the priority system introduced in Section 4.4.4 (in cases of hesitation, MOST has priority over ALL and SOME overrides the other two). Further, despite our efforts to make the task as simple as possible by encouraging paraphrasing, our annotation scheme still requires participants to compare the meaning of two separate statements (the original sentence and the explicitly quantified paraphrase). Schütze argues that such task demands too much of the annotators' memory. In further work, it would be beneficial to re-examine our guidelines and test variations thereof in order to get an even better understanding of the human process.

More directly, a future version of the scheme presented in this chapter should include guidelines regarding the selection of the referent of the noun phrase. It should also encourage the use of external resources to obtain the context of a given sentence, and give some pointers as to how to resolve issues or ambiguities caused by temporal quantification. In the meantime, we believe that the results as they stand are convincing enough to be taken as evidence that underquantification is analysable in a consistent way by humans. We also consider them as strong support for our claim that 'genericity quantifies'. In the next chapter, we start investigating whether the task can similarly be performed by machines.

Chapter 5

Automating quantification resolution

In Chapter 4, we showed that it was possible to manually quantify underspecified noun phrases with satisfactory agreement. This type of annotation, however, is expensive and cannot be applied in the areas of Natural Language Processing where large quantities of text are to be analysed. In this chapter, therefore, we turn to the problem of automating the annotation in a way that the correct logical form for an underquantified statement can be directly obtained from a syntactic or semantic parse. This requires not only selecting the correct quantifier for a noun phrase, but also obtaining its kind status and the distributive or collective status of the verbal predicate.

To our knowledge, no such automatic annotation has been attempted before. In consequence, we start our investigation with the simplest possible type of machine learning algorithm, using as determining features the direct syntactic context of the statement to be quantified. The general idea of such a system is that grammatical information such as the number of a subject noun phrase and the tense of its verbal predicate may be statistically related to its classification.

In what follows, we describe the creation of a gold standard for the task and give first insights into the distribution of quantifiers in underspecified sentences. We then present a basic statistical system in the form of a tree-based classifier which uses a small number of grammatical features. Our initial evaluation results, as reported in the last sections of this chapter, can be taken as a first baseline for the quantification resolution task.

5.1 The gold standard

As mentioned in Section 4.1.2, a machine learning system needs training data which, in some sense, exemplifies the task to be performed. The training data is assumed to be of high quality (that is, consistently annotated according to some guidelines) so that the system can learn general rules from it. In reference to the quality of the data, such a training set is usually called a **gold standard**.

5.1.1 Building the gold standard

In Chapter 4, we produced a corpus of 300 noun phrases annotated by three human subjects. In this section, we show how to use this annotation to produce a gold standard for the automatic classifier.

The simplest way to collapse the labels from the three annotations is to pick, for each noun phrase, the majority label. So if MOST appears twice out of three times for a particular instance, we mark that instance with the label MOST in the gold standard. The obvious limitation of that method is that, when the number of annotators is less or equal to the number of classes used in the annotation task, it is possible that no majority agreement exists for a given instance. Thus, another method, which avoids this pitfall, is to ask the annotators to negotiate and come to an agreement for each instance after the annotation task itself. The problem with this solution, however, is that the discussion tends to become over-reflective and undermines one of the assumptions of the annotation process: that it teaches us about the linguistic intuition of humans with regard to a particular classification (Teufel, to appear). Despite this known issue, negotiation is often used when lexicographic resources are produced.

In an attempt to produce the most accurate gold standard at a minimal cost, we combine both methods and use majority opinion when it is available and negotiation in cases of complete disagreement. Processing our results of Chapter 4, we find that majority agreement can be obtained for all cases of the distributive/collective annotation and the kind annotation. As for quantification, there are only 14 cases where a majority opinion cannot be obtained (see Section 4.5.2). This is a small overhead for our human subjects.

The main issue with the resulting gold standard is its relatively small size. The 300 data points it provides are clearly insufficient for machine learning purposes. On the other hand, the annotation process is time-consuming and we do not have the resources to set up a large-scale annotation effort. As a trade-off, we (the author of this thesis) annotated a further 300 noun phrases, thus doubling the size of the gold standard. The combination of the triple-annotated data and of the single-annotated, additional data was deemed acceptable as the author was shown to obtain overall substantial agreement with both other annotators in the experiments of Chapter 4. As an extra precaution, we also ran the classifier presented later in Section 5.2.3 over the original gold standard and over the author's annotations produced for Chapter 4 and found no substantial difference in performance between the two runs.

In what follows, we will use the phrase 'extended gold standard', or simply 'gold standard', to refer to the 600 data points obtained by merging the data of Chapter 4 and the additional 300 annotations.

5.1.2 Class distribution

Table 5.1 shows the class distribution of our five quantification labels over the 600 instances of the extended gold standard .

Class	Number of instances	Percentage of corpus
ONE	367	61%
SOME	53	9%
MOST	34	6%
ALL	102	17%
QUANT	44	7%

Table 5.1: Class distribution over 600 instances

We note, first, that the number of explicitly quantified noun phrases amounts to only 7% of the annotation set. This shows that the quantification task has potentially high value in the analysis of large amounts of data.

Next, we remark that 61% of all instances simply denote a single entity, leaving 32% to underquantified plurals — 189 instances. This imbalance is problematic for the machine learning task that we set out to achieve. First, it means that the training data available for SOME, MOST and ALL annotations is comparably sparse. Secondly, it implies that the baseline for our future classifier is relatively high: assuming a most frequent class baseline, we must beat 61% precision.

5.2 A syntax-based classifier

5.2.1 Some theory

Most of the remarks that can be found in the literature on the relation between syntax and quantification have been written with respect to the generic versus non-generic distinction. Although we have moved away from the terminology on genericity, we have gathered in this section a few examples that show the potential promises — and hurdles — of using syntax to induce quantification annotations.

We have already noted in Chapter 2 that there is no overt linguistic marker for genericity: if it is recognized that determiners, tense, aspect and number information do impact on the possible readings of a noun phrase as generic or non-generic, no firm rule can be established. Furthermore, it is rather the combination of various syntactic cues that gives information about the genericity of a noun phrase. The following gives an idea of which syntactic combinations can help identify genericity, and what the limits of such heuristics are.

The combination of a definite determiner *the* and a plural noun phrase usually blocks a generic reading (Chierchia, 1998):

225. The tigers ate meat.

However (Krifka et al, 1995):

226. The wolves are getting bigger as we travel north.

This feature is in fact irrelevant for the quantification task, as we showed in Section 3.3 that definite plurals could be interpreted as SOME, MOST and ALL statements.

Noun phrases which act as subjects of simple past tense verbs are usually non-generic (Gelman, 2004):

227. A cow says 'moo'.

228. A cow said 'moo'.

However, the so-called 'historic past' is an exception to this rule:

229. The woolly mammoth roamed the earth many years ago.

The combination of a bare plural and present tense is a prototypical indication of genericity (Cimpian and Markman, 2008):

230. Tigers are massive.

Although not in news headlines:

231. Cambridge students steal cow.

Table 5.2 shows the distribution of various grammatical constructions with respect to quantification, as obtained from our gold standard. The percentages shown correspond to the ratio of each annotation label for a particular construction. So for instance, 45% of definite plurals followed by a past tense can be annotated as ALL.

Although some constructions give a clear majority to one or another label, that majority is not always overwhelming. For instance, consistently annotating bare plurals followed by a past tense as SOME would result in a precision of only 54%. It is therefore unclear how accurate a classifier based only on syntax can be. This is what we investigate in the rest of this chapter.

Construction	ONE	SOME	MOST	ALL
bare plural	6%	40%	13%	35%
bare plural + present	0%	22%	17%	56%
bare plural + past tense	17%	54%	6%	17%
definite singular	94%	0%	1%	4%
definite singular + present	91%	0%	2%	5%
definite singular + past tense	97%	1%	0%	2%
definite plural	4%	11%	29%	53%
definite plural + present	0%	0%	29%	71%
definite plural + past tense	10%	15%	25%	45%
indefinite singular	68%	5%	5%	21%
indefinite singular + present	52%	10%	5%	33%
indefinite singular + past tense	100%	0%	0%	0%

Table 5.2: Corpus Statistics

5.2.2 Features

We choose classification features with respect to the observations previously made on genericity (see Section 5.2.1). We give the system article and number information for the noun phrase under consideration, as well as the tense of the verbal predicate following it. In order to cater for proper nouns, we also indicate whether the head of the noun phrase is capitalised or not. Article, number and capitalisation information is provided for the object of the verb. All features are automatically extracted from the RMRS presentation of the sentence in which the noun phrase appears (see Section 4.2.1 for details). The following shows an example of a feature line for a particular noun phrase:

TRIPLE: influence include artist

ORIGINAL: His early blues influences included artists such as Robert Johnson, Bukka White, Skip James, Jerry Miller and Sleepy John Estes.

FEATURES: past,possessive,plural,nocap,bare,plural,nocap

Note that articles belonging to the same class are labelled according to the class: all possessive articles, for instance, are simply marked as 'possessive'. This is the same for demonstrative articles.

We expect that such syntactic information will help with the task of quantification. It is less clear whether this is at all useful for deciding of the kind of value of the noun phrase or making the distinction between collective and distributive predicates. We expect, in particular, that the latter would be bound to the lexical semantics of the verb. We however report how our simple syntactic system deals with those classification tasks and use the

results as a baseline for further investigations in Chapter 6.

It must also be noted that the parse obtained from the RASP/RMRS pipeline is not perfect. We will come back in the discussion to the problems posed by incorrect, or incomplete, parses.

5.2.3 The classifier

The aim of this work is not only to produce an automatic quantification system, but also, if possible, to learn about the linguistic phenomena surrounding the underspecification of quantification. Because of this, we choose a tree-based classifier which has the advantage of letting us see the rules that are created by the system and thereby may allow us to make some linguistic observations with regard to the cooccurrence of certain quantification classes with certain grammatical constructions.

In this work, we use an off-the-shelf implementation of the C4.5 classifier (Quinlan, 1993) included in the Weka data mining software.[1] We give next an overview of how C4.5 produces decision trees based on some annotated data.

The C4.5 classifier

The C4.5 algorithm functions in three simple steps:

> For each feature F in the data, find the (normalised) information gain given by F
>
> Given F_{BEST}, the feature with the highest normalised information gain, create a decision node that splits the decision tree on F_{BEST}
>
> Recurse on each branch of the decision tree separately

The information of a given probability distribution P is defined as:

$$I(P) = \sum_{i=1}^{k} p_i \quad log(p_i) \tag{5.1}$$

where k is the number of classes in the distribution (in the case of the quantification annotation, 5) and p_i is simply the probability of a given class: for instance, given the results given by Table 5.1, $p_{one} = 0.61$ in the first iteration of the algorithm.

Further, the information of the distribution after partitioning the data according to feature F is:

[1] http://www.cs.waikato.ac.nz/ml/weka/

$$I(P,F) = \sum_{i}^{n} \frac{P_i}{P} I(P_i) \tag{5.2}$$

where $P_{1...n}$ are the partitions created by splitting the data on feature F (n is the number of values that F can take). In our case, assuming a split on the number feature which can take two values, singular or plural, n is equal to 2, P is the number of instances in the data set (600 at first iteration), P_{sg} is the number of plural instances and P_{pl} the number of singular instances. $I(P_i)$ is calculated as before on the partitions of all singular instances and all plural instances respectively.

The information gain for a particular feature is the difference between the information of the entire distribution and the information of the distribution after partitioning:

$$Gain(P,F) = I(P) \quad I(P,F) \tag{5.3}$$

5.2.4 Experimental setup

We run the classifier separately on the quantification data, the kind annotations and the distributive/collective labels. For each type of data, we perform a 6-fold cross-validation on the gold standard and report precision, recall and F-score for each class. Precision is taken as the number of correctly classified instances for a particular class over the number of instances returned by the system for that class. Recall is calculated as the number of correctly classified instances for a class over the actual number of instances of that class in the test set. The F-score is the weighted harmonic mean of precision and recall, as given by:

$$F = \frac{2 \quad precision \quad recall}{precision + recall} \tag{5.4}$$

5.3 Results and discussion

5.3.1 Results

The C4.5 classifier gives 78%, 89% and 85% overall precision to the quantification task, the kind annotation and the distributive/collective annotation, respectively. Tables 5.3 to 5.5 show per class results for the three tasks. The figures in brackets indicate the number of true positives for a particular class, followed by the total number of instances annotated by the system as instances of that class. As far as quantification is concerned, the classifier performs extremely well with the ONE class, reaching 92% F-score. Already quantified noun phrases, as expected, yield perfect precision and mediocre recall. The

Class	Precision	Recall	F-score
ONE	86% (362/422)	99% (362/367)	92%
SOME	60% (25/42)	47% (25/53)	53%
MOST	33% (2/6)	6% (2/34)	10%
ALL	53% (57/108)	56% (57/102)	54%
QUANT	100% (22/22)	50% (22/44)	67%

Table 5.3: Class precision and recall for the quantification task

Class	Precision	Recall	F-score
KIND	–	0% (0/54)	–
NOTKIND	87% (502/575)	100% (502/502)	93%
QUANT	100% (25/25)	57% (25/44)	73%

Table 5.4: Class precision and recall for the kind annotation

Class	Precision	Recall	F-score
DIST	87% (452/519)	95% (452/477)	91%
COLL	51% (30/59)	38% (30/79)	44%
QUANT	100% (22/22)	50% (22/44)	67%

Table 5.5: Class precision and recall for the distributive/collective distinction

system also seems to mirror the behaviour of human annotators by performing less well with the labels SOME, MOST and ALL.

In order to understand the distribution of errors, we perform a detailed analysis on the first fold of our data. Out of 100 instances, the classifier assigns 25 to an incorrect class. The majority of those errors (44%) are due to the fact that the classifier labels all singulars as ONE (see Output 5 in the next section), missing out on generic interpretations and in particular on the plural reading of mass terms: out of 11 errors, 5 are linked to a bare singular). The next most frequent type of error, covering another 16% of incorrectly classified instances, comes from already quantified noun phrases being labelled as another class. These errors affect the recall of the QUANT class, as discussed further below, and the precision of the SOME, MOST and ALL labels in particular (most of those errors occur in plural noun phrases). The coarseness of the rules is again to blame for the remaining errors: looking at the decision tree produced by the classifier, we observe that all bare plurals followed by a present tense, as well as all definite plurals, are labelled as universals, while all bare plurals followed by a past tense are labelled as SOME. This accounts for a further 7 errors. The last three incorrect assignments are due to a dubious capitalisation rule.

A partial answer to the recall issue with regard to the QUANT label would be to provide the system with a comprehensive list of quantifiers. This, unfortunately, would not produce perfect recall as the other problem we encounter is linked to the parsing pipeline: in some cases, the article of the noun phrase is shown as a modifier in the syntactic parse, leading to a semantic representation which is not what we expect from a generalised quantifier. A full solution would therefore involve some work on the parser or a post-parsing step in which the system looks at the surface form of all modifiers and attempts to decide whether they are actually quantifiers rather than adjectives. Despite the need for a solution to this problem, we will not discuss it further in this thesis as it purely concerns implementation rather than quantification theory.

Surprisingly, the grammatical features contribute to some extent to the elucidation of the distributive/collective distinction. Distributives attain 87% precision while collectives are classified with 51% precision, which, although relatively low, is also higher than expected. Looking at the decision trees produced by the classifier, we learn that the higher precision obtained for the distributive class is due in great part to a simple rule that labels all singulars as DIST. Given that most singular noun phrases refer to individuals rather than groups, the rule is appropriate from a statistical point of view.[2] The rules that successfully classify collective statements focus on the combination of the article of the noun phrase and the tense of the verb. For instance, a definite plural followed by a past tense tends to be collective. It is difficult to find a particular linguistic reason for such a preference, so we will assume that it is produced by the particular distribution of our corpus and not necessarily generalisable. Collectives also benefit from the fact that the two other classes are classified with high F-score.

As expected, the algorithm is unable to deal with the kind distinction and ends up classifying every (non-quantified) instance as NOTKIND.

5.3.2 Some correlations with linguistic theory

Output 5 shows one of the decision trees produced by the classifier for the quantification task (Fold 1). We observe that most definite plurals (including demonstratives and possessives) are classified as either MOST or ALL. This fits the linguistic notion of a definite as being essentially universal (see reference to Lyons in Section 2.3) but also misses out on the correct quantification of statements such as 225:

225. The reporters asked questions after the press conference.

We note also that non-capitalised bare plurals followed by a present tense are similarly classed as ALL. This echoes the observation that the combination of bare plural and

[2] Although a favourite of the linguistic literature, statements of the form *The duck lays eggs* are actually relatively rare in normal text.

Output 5 Tree output of the C4.5 classifier in Weka. Training on Folds 2-6. Quantification task.

```
number1 = sg: one (345.0/37.0)
number1 = pl
|   article1 = a: all (0.0)
|   article1 = the: all (40.0/17.0)
|   article1 = dem: all (4.0)
|   article1 = poss
|   |   tense = present
|   |   |   number2 = sg: all (2.0)
|   |   |   number2 = pl: most (2.0)
|   |   tense = past: all (6.0/1.0)
|   |   tense = perfect: all (0.0)
|   |   tense = future: all (0.0)
|   |   tense = ing: most (1.0)
|   |   tense = progressive: all (0.0)
|   article1 = card: quantified (4.0)
|   article1 = one: all (0.0)
|   article1 = both: quantified (1.0)
|   article1 = some: quantified (10.0)
|   article1 = several: quantified (1.0)
|   article1 = most: quantified (3.0)
|   article1 = many: quantified (5.0)
|   article1 = any: all (0.0)
|   article1 = each: all (0.0)
|   article1 = no: all (0.0)
|   article1 = another: all (0.0)
|   article1 = which: all (0.0)
|   article1 = every: all (0.0)
|   article1 = all: quantified (2.0)
|   article1 = null
|   |   capleft = cap: one (6.0/1.0)
|   |   capleft = low
|   |   |   tense = present: all (35.0/17.0)
|   |   |   tense = past: some (23.0/9.0)
|   |   |   tense = perfect: some (8.0/2.0)
|   |   |   tense = future: some (0.0)
|   |   |   tense = ing: all (1.0)
|   |   |   tense = progressive: all (1.0)

Number of Leaves  :     32
Size of the tree :      38
```

present is a typical manifestation of genericity (if one understands genericity as a quantification phenomenon close to universality — see the inductivist approaches summarised in Section 2.2.4). When followed by past or perfect tenses, an existential quantification with SOME is however preferred.

One of the puzzles of the decision tree is the use of the 'number2' feature to distinguish between MOST and ALL in the case of some definite plurals (training on other folds actually shows similar tendencies). The 'number2' feature corresponds to the number of the object noun phrase in the sentence. So given Sentences 226 and 227, a system using the classifier in Output 5 would label the first one as ALL and the second one as MOST.

226. *My cats* like the armchair. ALL

227. *My cats* like the armchairs. MOST

At first glance, the rule seems to be a mere statistical effect of our data. We will however remark that statements like 227 are reserved a special section in Link (1998), where they are introduced as 'relational plural sentences'. One of Link's claims is that those sentences warrant four collective/distributive combinations — as opposed to two only in the case where the object is an individual. So we can say in Sentence 227 that a collective of cats likes a collective of armchairs, or that this collective of cats likes each armchair individually, etc. This proliferation of interpretations makes uncertainties more likely with regard to who likes what, and to the quantification of the subject and object. We will finally remark with Link that lexical semantics play a role in the interpretation of the collective/distributive distinction (see how many combinations are plausible in 228) and that, if this distinction is actually correlated with a particular distribution of quantifiers, then lexical semantics plays a role in quantification resolution. This fits the natural expectation that a classifier with more semantic knowledge would be more able to correctly distinguish between the various partitions of the quantificational space.

228. The women released the prisoners. (Link, 1998)

For now, we will simply conclude that, although a simple syntax-based classifier is able to classify certain constructs with high precision, other constructs are beyond its capabilities. Further, it is difficult to see how improvements can be made to the current classification without venturing outside of the grammatical context. For instance, it seems practically impossible to improve on the high-precision rule specifying that every singular noun phrase should be classified as ONE. Accordingly, the next chapter investigates the use of lexical semantics to break those limitations.

Chapter 6

Quantifying with similarity

We have shown so far that syntax alone is not sufficiently informative to let us automate quantification resolution. In this chapter, we will first discuss the locus of quantification and ask what level of semantics, or indeed pragmatics, is needed to achieve our aim. We will argue that in the majority of cases, quantification is situated beyond semantics but that methods based on semantics can help us get access to the world knowledge that humans use when interpreting underquantified statements.

We will then show how problem-solving skills, as derived from human psychology, are relevant to quantification resolution. We will focus on memory-based reasoning (Stanfill and Waltz, 1986) — the use of analogy to process new stimuli — as a tool to access the pragmatics behind underquantified statements. From an implementation point of view, we suggest the use of distributional similarity as a way to compute analogy between the statements to be quantified and 'memorised' statements.

We finish this chapter, as well as the core of this thesis, with results of a quantification system based on memory-based learning.

6.1 Where is the quantification constraint?

There is nothing wrong, syntactically, in saying that *some cats are mammals* or that *most ducks lay eggs*. No more, in fact, than uttering the following:

229. My toothbrush is alive and trying to kill me. (Jacobs, 1969).

The problem with 229 is one of selectional restriction, and the hearer, before referring the speaker to a psychiatry clinic as suggested by McCawley (1971), might point out that toothbrushes are objects and therefore not supposed to be alive or to partake in activities associated with sentient beings. Similarly, it is possible to argue that the property of being

a mammal, in virtue of its semantics, can only ever apply to *all* members of a species. We could thus hypothesise that quantification is in some sort a problem of selectional restriction and encode appropriate constraints in our parser's lexicon (assuming a deep grammar is available). This would situate quantification at the level of compositional semantics. Note that such a lexical effort is in no way trivial to achieve. In our example involving the predicate *mammal'*, the restriction must be encoded at the level of the entire VP and must access the determiner of the noun phrase that the VP eventually selects as its subject. It is much more complex than the encoding of selection restriction as it is normally understood (i.e. as the properties of an entity itself, not of its quantification).

One further problem with this hypothesis is that the constraints that can in theory be implemented do not all stem from strict selectional restriction, as required by the lexical semantics of the individual components of the verb phrase, but also from pure world knowledge. In the latter cases, it is probably preferable to talk of selectional preference. For instance, the predicate *to discover a lake* can in theory apply to all humans (and relevant extraterrestrials) but in practice, only a small number of those are concerned.

It must be remarked, finally, that the quantification resolution of many statements depends not even on matters of selectional preference but purely on the context of the sentence. It is for instance highly characteristic of a journalist to ask questions but in Sentence 230, the context of the press conference makes it likely that only a small number of journalists will have asked questions.

230. The journalists asked the President questions at the end of the press conference.

In those cases, it is likely that the mechanism employed by humans to resolve the referent's quantification rather involves knowledge of the scenario under consideration: we know what a press conference is like, who takes part in it and in what way. That is, we recognise the **situation** and make correct inferences given our knowledge of it.

We want to suggest that, although lexical semantics may be directly at work in examples where a rules and regulations interpretation is the only one possible (leading to a universal reading of the subject noun phrase — see Sentences 231 and 232), we must step beyond semantics to resolve most other cases.

231. Cats are mammals.

232. In chess, bishops move diagonally.

In the rest of this chapter, we will attempt to reproduce some of the pragmatic inferences that are made by humans when resolving quantification. Because the focus of this thesis is quantification itself, and not the related phenomena of distributivity (against collectivity) and kind interpretation, we will endeavour to create a system implementation that

responds to this focus. We will, however, make the simplifying assumption that the pragmatics involved may be of use when considering the other annotations needed for our proposed formalisation and explore the validity of this assumption.

6.2 Situational analogy

We have made the hypothesis that quantification resolution involves problem-solving at the pragmatic level. In the rest of this work, we will investigate analogy as a potential way to find out the quantification value of a particular noun phrase. We refer to Schank and Abelson's theory of episodic memory (1977) and their introduction of **scripts** to explain pragmatic processing of commonly experienced situations. Schank and Abelson proposed that memory is organised via prototypes of autobiographical events. Those allow humans to make inferences about untold information in new narratives. So when we read Sentence 233, we assume that Mary paid *the bill* at the restaurant because we know that going to the restaurant involves getting food and having to settle a bill at the end of the meal.

233. We went to the restaurant last night. Mary paid.

We suggest that quantification also can be understood through reference to memorised scripts, or situations, where the interaction between certain kinds of entities implies (mostly) fixed quantities.

For instance, the sentence *The child broke the teapot* calls up a particular script involving an individual breaking a piece of crockery. This script, although allowing various interpretations, depending on the syntactic environment of the statement, blocks a wide range of readings. The following shows some possible and some blocked interpretations for three syntactic environments:

234. Children broke the teapot.
    (a) Some children, collectively, broke the teapot.
    (b) *All children, in those times, regularly broke a particular teapot.

235. The child broke the teapot.
    (a) One child broke a particular teapot.
    (b) *Some children, in those times, regularly broke a particular teapot.

236. The children broke the teapot.
    (a) The children, collectively, broke a particular teapot.
    (b) *Most of the children, one after the other, broke the teapot.

Note that regardless of the person and the piece of crockery involved, the blocking effect remains: *The shop assistant broke a cup.* Conversely, moving to a different script (that is, to a statement involving the composition of lexically dissimilar entities) changes the range of readings available:

237. Members of the government broke their promises.

    (a) Some/Most/All members of the government, each individually, broke their promises.

We propose to simplify the definition of a script as an event of the type 'X does E' where the quantification of X is known. The problem-solving task becomes one of identifying the most likely script for a new event. To do this, we compute the **situational analogy** between the new event and all previously 'experienced' events available in our annotated **script database**: given a situation S_1 where some entity X_1 is engaging in some event E_1 and where the quantification of the subject Q_1 is unknown, if we can find a script S_2 where an entity X_2, similar to X_1, engages in some event E_2, similar to E_2, for which we know the quantification Q_2 of the subject, then $Q_1 = Q_2$. We can thus implement a classification system relying on this hypothesis by trying to find, in an annotated set, the statement most similar to the one to be quantified.

The idea that analogy can help solve complex pragmatics-related tasks is not novel. Stanfill and Waltz (1986) first advocated the use of 'memory-based reasoning' (that is, k nearest neighbours search) in artificial intelligence as a way to overcome the issues encountered by rule-based paradigms. Their aim was to build systems closer to what they thought 'real time' human reasoning was like, and to avoid the spurious correlations identified by rule-based algorithms as relevant to a given classification task. Daelemans et al (1999) showed later that ignoring outliers in training data could actually affect negatively the results of a range of natural language processing tasks and suggested memory-based reasoning as a suitable solution to this problem.

In the next subsection, we review various definitions of similarity and argue for the distributional view as the adequate tool for our task, given the various technical constraints that affect the implementation.

6.2.1 Which similarity?

As we have just seen, a good script for a new event will be one where the interacting entities resemble those in that event. But although orcs are creatures like humans, rings are objects like teapots, and the verbs *destroy* and *break* are near-synonyms, the annotations for *The children broke the teapot* and *The orcs destroyed the ring* may be different (the

latter may refer to the orcs as a people — leading to the annotation SOME). So the question that we face is, what kind of similarity is appropriate when comparing situations?

In what follows, we consider several ways to compute similarity and hypothesise their effect on a quantification resolution system, both in terms of accuracy and coverage. We refer to **accuracy** as the amount of correct annotations performed by the system and to **coverage** as the amount of annotations that the system can perform at all. (We posit that the annotation should only be performed if the best script is sufficiently similar to the event under consideration). We assume that, when using a fairly loose notion of similarity, we will be able to annotate more data given our training set — but with possibly less accuracy, while using a stringent notion of similarity should improve accuracy but decrease coverage (because a matching script will not necessarily be found for each new event).

Traditionally, lexical similarity has been computed using **distributional similarity** methods. The assumption of distributional similarity (Harris, 1954) states that two words appearing in similar contexts will be close in meaning. This observation is statistically useful and has contributed to successful NLP systems dedicated to the extraction of words similar to an input seed. Those systems can be classed in two approaches:

> The pattern-based approach (e.g. Ravichadran and Hovy, 2002). The most significant contexts for the input seed are extracted as features and those features used to discover words related to the input (under the assumption that words appearing in *at least one* significant context are similar to the seed word). There is also a non-distributional strand of this approach: it uses Hearst-like patterns (Hearst, 1992) which are supposed to indicate the presence of two terms in a certain relation — most often hyponymy or meronymy (see Chklovski and Pantel, 2004).
>
> The feature vector approach (e.g. Lin and Pantel, 2001). This method fully embraces the definition of distributional similarity by making the assumption that two words appearing in similar *sets* of features must be related.

In distributional similarity methods, the features found for a word come from general text, and as such are representative of the discourse surrounding that word. We will thus call them **discursive features**. Discursive features are normally not particularly enlightening as to what the definition of that word is, that is, they do not reflect what the encyclopaedic article for that word would deem important. (Even when an encyclopaedia forms the corpus into consideration, distributional similarity is calculated over the whole corpus — in order to get enough data — and not over a single article.) For instance, the 30 most characteristic contexts found by our own system (see baseline in Herbelot, 2009) in a 500MB subset of Wikipedia for the word *brother* include:

brother – inherit – throne

brother – split – inheritance

brother – show – symptom.

Some of the returned features may be considered as discursively prominent in an encyclopaedia but they are in no way indicative of what a brother is. There is also no ground to believe that a more general corpus would change the situation much. The fact is that we very seldom say 'a brother, in relation to another person, is an individual of male sex who shares parents with that person'. As a result, when we try to compute the terms most similar to *brother*, we obtain a list starting with *son, descendant, father, grandson* and going on to *leader, child, duchess, kingdom*, etc. The first group is to some extent homogeneous: it is possible to make all terms hyponyms of the concept 'family member'. The rest of the list, however, is made of loosely related items which only share a general feature with the original word: leaders, children and duchesses are humans like brothers. *Kingdom* is in no way similar to the seed and can be taken as a system error.

We expect then that distributional similarity would help us where a loose notion of similarity is needed, yielding high coverage and lower accuracy.

The stringent version of similarity is **entailment**, as proposed by Geffet and Dagan (2005). Entailment corresponds to a 'vertical' notion of similarity — something close to the hyponymic relationship where one entity is the parent of another one in a taxonomy (for a discussion of the relation between entailment and hyponymy, see Croft and Cruse, 2004). A typical test for identifying whether two words are in an entailment relation is to find contexts where one of the words is substitutable for the other without any change of meaning (Szpektor et al, 2007). Geffet and Dagan (2005) make the assumption that if word w entails word y, then w's set of features should include all features contained in y's set. So for instance, *cat* and *feline* are in an entailment relation because everything that could be said of a feline could be said of a cat as well. They actually compute entailment relations using distributional similarity and therefore rely on discursive features. It is arguable whether this type of feature is best for the computation, or whether intrinsic features should be used, i.e. features that reflect what the concept is rather than how it is used. Regardless, when applying entailment to our genericity system, we can expect high accuracy but rather low coverage (unless our script database is very large).

We will call the last type of similarity **horizontal** in reference to a taxonomic structure where all children of a given node are on the same level. This similarity is halfway between distributional similarity and entailment, in that two items are not required to have feature sets in an inclusion relation (as in the cat – feline example) but the amount of intrinsic features that they share is expected to be large (as opposed to the brother – duchess example which mostly shares discursive features but few intrinsic ones). This may be

the best compromise in terms of precision and recall for situational analogy, but relies on the encyclopaedic descriptions of the terms under consideration (in order to find out the intrinsic features of the concepts) and cannot be performed using arbitrary text.

Our choice is motivated by several technical constraints. First, entailment requires a very large annotated database which we do not have at our disposal. Secondly, horizontal similarity relies on dictionary or encyclopaedic descriptions which are both comprehensive (a sufficient number of features must be present to calculate similarity at all) and focused (we are only interested in the intrinsic features of a concept, not in side information). Dictionaries fail on the former requirements while encyclopaedias fail on the latter. Further, encyclopaedias do not cater for verbs at all and both types of resources may have an incomplete coverage of nouns. We therefore settle on the loosest notion of similarity. In the following section, we present an implementation of an annotation system based on situational analogy with distributional similarity.

6.3 Situational analogy: system implementation

6.3.1 The data

We reuse the gold standard that we produced in Chapter 5. We perform 6-fold cross validation on the data, creating for each fold a knowledge base of 500 annotated statements, which will be used as **reference scripts** by the algorithm, and a test set of 100 statements on which we will calculate the precision of our system.

6.3.2 The similarity measure

Although an ideal system would use as large a context as possible for the noun phrase to be annotated, we prefer the robustness of a simpler system and start with the core of the statement itself, i.e. its head words.

We make the assumption that the similarity of two situations S and R can be linearly computed from the similarity of their respective components:

$$sim_{triple}((S_1, R_1), (S_2, R_2), (S_3, R_3)) = sim(S_1, R_1) + sim(S_2, R_2) + sim(S_3, R_3)$$

That is, given two statements consisting of a subject, a verb and an object, we compute the similarities of both subject, both verb and both object heads separately and sum those to obtain the similarity for the whole statements. An example is given in Output 6.

In what follows, we introduce the technique we use for calculating distributional similarity over pairs of words.

Output 6 Example output

```
**************************************************
Situation 1: father give crown
Situation 2: son inherit task
**************************************************
S -- father son -- Similarity: 0.136053
V -- give inherit -- Similarity: 0.00368466
O -- crown task -- Similarity: 0.0432487
**************************************************
Overall Similarity: 0.182986
**************************************************
```

The background corpus

The corpus used for our distributional similarity baseline consists of a subset of Wikipedia totalling 500 MB in size, parsed first with RASP2 (Briscoe et al, 2006) and then into a Robust Minimal Recursion Semantics form (RMRS, Copestake, 2004) using the RASP to RMRS converter (Ritchie, 2004). We have already described in Section 4.2.1 the structure of an RMRS output. We will now come back to the description of the flat representation obtainable from the parse and show how it can be used as basis for distributional methods.

The RMRS representation consists of trees (or tree fragments when a complete parse is not possible) which comprise, for each phrase in the sentence, a semantic head and its arguments. For instance, in the sentence *Owls lay white eggs*, three subtrees can be extracted:

`lemma:lay arg:ARG1 var:owl`

which indicates that *owl* is subject of the head *lay*,

`lemma:lay arg:ARG2 var:egg`

which indicates that *egg* is object of the head *egg*, and

`lemma:white arg:ARG1 var:egg`

which indicates that the argument of *white* is *egg*.

Note that any tree can be transformed into a discursive feature for a particular lexical item by replacing the slot containing the word with a hole: `lemma:lay arg:ARG2 var:egg` becomes `lemma:lay arg:ARG2 var:hole_`, a potentially characteristic context for *egg*.

Given the nature of our data, we only need similarity figures for nouns and transitive verbs. In order to speed up processing, we reduce the RMRS corpus to two subcorpora,

one for each part of speech. The first subcorpus consists of all relations including nouns. The second subcorpus consists of a list of relations with a verbal head and at least two arguments: `lemma:verb-query arg:ARG1 var:subject arg:ARG2 var:object`. Note that we do not force noun phrases in the second argument of the relations and for instance, the verb *say* is both considered as taking a noun or a clause as second argument (*to say a word*, *to say that...*).

Algorithm

The similarity algorithm relies on the idea that two words that are similar will have similar feature vectors (see Geffet and Dagan, 2005). We define here the feature vector of word w as the list of discursive features containing w, together with the Pointwise Mutual Information (PMI) of each feature in relation to w as a weight. PMI is defined as follows:

$$pmi(f, w) = \log\left(\frac{P(f, w)}{P(f)\,P(w)}\right) \tag{6.1}$$

where $P(f)$ and $P(w)$ are the probabilities of occurrence of the feature and the word respectively and $P(f, w)$ is the probability that they appear together.

PMI is known to have a bias towards less frequent events. In order to counterbalance that bias, we apply a simple logarithm function to the results as a discount (we multiply the original PMI value by this discount to find the final PMI):

$$d = \log(c_{wf} + 1) \tag{6.2}$$

where c_{wf} is the cooccurrence count of a word and a feature. This function provides an actual discount for cooccurrences observed only once in the corpus (those single relations between a word and a feature may be parsing errors or simply odd, non-informative combinations). For cooccurrences that are observed more than once, the discount has the effect of giving a larger increase to the PMI of sufficiently frequent events and a smaller, more marginal increase to the PMI of events that are only observed a few times. For instance, a PMI where $c_{wf} = 2$ would be multiplied by 1.1 while in a case where $c_{wf} = 10$, the PMI would receive an increase by a factor of 2.4.

We compared the proposed discount with that suggested in Pantel and Ravichandran (2004):

$$d = \frac{c_{wf}}{c_{wf} + 1} \quad \frac{min\left(\sum_{i=1}^{M} c_{wi}, \sum_{j=1}^{N} c_{fj}\right)}{min\left(\sum_{i=1}^{M} c_{wi}, \sum_{j=1}^{N} c_{fj}\right) + 1} \tag{6.3}$$

where c_{wf} is the cooccurrence count of an instance and a feature, M the number of words and N the number of features in the corpus.

Tested on the experiments reported at the end of this section, the two discount factors give identical results. In the rest of this thesis, we use our own function.

For each pair of words $(w1, w2)$ we extract the feature vectors of both $w1$ and $w2$ and calculate their similarity using the measure of Lin (1998):

$$Lin(w_1, w_2) = \frac{\sum_{f \in F_{w_1} \cap F_{w_2}} [W(f, w_1) + W(f, w_2)]}{\sum_{f \in F_{w_1}} W(f, w_1) + \sum_{f \in F_{w_2}} W(f, w_2)} \tag{6.4}$$

where F_w is the feature vector for word w and $W(f, w)$ is the weight of feature f for word w (in our system, the corresponding PMI).

As a check of how the Lin measure performed on our Wikipedia subset using RMRS features, we reproduced the Miller and Charles experiment (1991) which consists in asking humans to rate the similarity of 30 noun pairs. The experiment is a standard test for semantic similarity systems (see Jarmasz and Szpakowicz, 2003; Lin, 1998; Resnik, 1995 and Hirst and St Onge, 1998 amongst others). The correlations obtained by previous systems range between just below 0.7 and just below 0.9. Those systems rely on edge counting using manually-created resources such as WordNet and the Roget's Thesaurus. Given enough data, similar performance can be obtained automatically: Bollegala et al (2007) report a correlation over 0.8 using the web as data set and a method involving Google hits and automatically extracted patterns indicative of synonymy.

Applying our feature vector step to the Miller and Charles pairs, we get a correlation of 0.38, way below the edge-counting systems. It turns out, however, that this low result is at least partially due to data sparsity: when ignoring the pairs containing at least one word with frequency under 200 (8 of them, which means ending up with 22 pairs left out of the initial 30), the correlation goes up to 0.69. This is in line with the edge-counting systems and shows that our baseline system produces a decent approximation of human performance, as long as enough data is supplied.

Having implemented the distributional similarity algorithm and tested it on the Miller and Charles set, we proceed to calculate similarities for all pairs of nouns and all pairs of verbs in our gold standard. With a count of 600 statements comprising one verb and two nouns each, we store a database of 1200^2 noun similarities and 600^2 verb similarities.

In order to avoid putting too much weight on statements that share one or more words (because of the identical words, the overall similarity of the statements can jump over 1 or 2 when in fact, a 'good' similarity might be anything over 0.1), we record the second highest similarity for all noun pairs and verb pairs and convert all similarities of 1 to the appropriate value: in our system, 0.26 for nouns and 0.04 for verbs.

Finally, we normalise all scores to get a similarity value between 0 and 1.

6.3.3 The situational analogy algorithm

We implement a nearest neighbour algorithm, as it is the most direct implementation of memory-based learning.

For each noun phrase X to quantify in a statement S_0, we calculate the similarity between S_0 and every script $S_{1...n}$ in our database. We then pick the script with the highest similarity score and copy the quantification of its subject noun phrase.

This algorithm can be modified in several ways. First, we argued in Section 6.2 that the syntax in the statement had an influence on the potential readings. So it may be beneficial to only consider the scripts that share the same syntactic context as the statement to be annotated. Doing so may have a positive impact on precision but will inevitably affect recall. In Section 6.4, we report results of experiments that use various amounts of syntactic information.

Secondly, we also mentioned that annotation should only be performed when the top similarity for a given statement is sufficiently high. To achieve this, we can set up the system so that it falls back on the rules extracted by the statistical classifier in Chapter 5 when the similarity score is too low. Various thresholds can be applied and we report the results of various experiments in Section 6.4.

6.4 Results

Table 6.1 repeats the overall precisions obtained by the C4.5 classifier in Chapter 5 for the three subtasks of quantification, distributive against collective classification and kind annotation.

	Quantification	Distributive/Collective	Kind
Precision of rules	78%	85%	89%

Table 6.1: Overall precision for all subtasks, as obtained from the C4.5 classifier

Our results for the nearest neighbour algorithm illustrate two trends. First, as expected, the similarity module functions better when more syntactic context is provided to the system. Tables 6.2 to 6.4 show the results of experiments performed for our three subtasks, with varying degrees of syntactic information used. We assess the situational analogy algorithm on its own, that is, we do not fall back on C4.5 rules when classification is impossible. We report accuracy and coverage for the system when no syntax is provided, with article and number information for the subject noun phrase only, with tense added, and finally with the full syntactic context (including the article and number of the object noun phrase). Class accuracy is also given.

We observe that the more syntax is available, the more the accuracy increases (with an expected loss of coverage). The best accuracy is obtained in all three tasks when the full context is provided to the system. All further experiments reported in this chapter accordingly use all five syntactic features for classification. We use strict matching of context.

The second trend is that in the quantification task and the kind annotation task, as well as (to a lesser extent) in the case of the distributive/collective distinction, the precision of the situational analogy module increases when we only consider scripts with a score above a certain threshold. Tables 6.5 to 6.7 show the results of experiments performed with different thresholds. The column entitled 'no rules' reports precision and recall when the C4.5 rules are fully ignored (when the maximum similarity reported by the system is 0, we default to the most frequent class for the task). We note that for quantification and kind, a constant increase in precision can be observed as the threshold goes up (see the 'precision of situational analogy' row). A similar situation is illustrated by the figures for the distributive/collective subtask, although a drop is seen at the highest threshold. Overall precision (including the annotations given by the rules) shows a steady increase for all three tasks.

	No syntax	art1, num1	art1, num1, tense	full syntax
Accuracy	51%	69%	70%	72%
Coverage	100%	98%	95%	84%
ONE	70% (250/355)	89% (308/346)	88% (310/353)	88% (297/339)
SOME	23% (10/44)	33% (17/51)	36% (16/44)	32% (12/38)
MOST	8% (2/25)	14% (4/28)	17% (4/23)	38% (10/26)
ALL	31% (39/124)	41% (54/133)	42% (52/124)	38% (33/88)
QUANT	12% (6/50)	80% (24/30)	75% (21/28)	71% (10/14)

Table 6.2: The effect of syntax. Results for the quantification annotation.

	No syntax	art1, num1	art1, num1, tense	full syntax
Accuracy	74%	81%	84%	85%
Coverage	100%	98%	95%	84%
NOTKIND	86% (427/498)	90% (441/492)	91% (444/489)	90% (409/454)
KIND	25% (13/52)	15% (10/66)	24% (13/55)	27% (10/37)
QUANT	12% (6/50)	80% (24/30)	75% (21/28)	71% (10/14)

Table 6.3: The effect of syntax. Results for the kind annotation.

In order to ascertain the extent to which the threshold positively influences the results, we produce learning curves for the three subtasks: we collect all the similarities used for

	No syntax	art1, num1	art1, num1, tense	full syntax
Accuracy	65%	81%	81%	83%
Coverage	100%	93%	95%	84%
DIST	81% (369/456)	90% (418/465)	91% (408/446)	91% (382/419)
COLL	18% (17/94)	40% (37/93)	38% (37/98)	40% (29/72)
QUANT	12% (6/50)	80% (24/30)	75% (21/28)	71% (10/14)

Table 6.4: The effect of syntax. Results for the distributive/collective distinction.

annotation when the threshold is 0, $sim_{1...600}$, and sort them by increasing value. We then calculate the precision of the system over all 600 items, then over the 599 items with highest similarities, then over 598, etc, so that we get a precision figure at each recall point. The resulting precision curves are shown in Figure 6.1.

Quantification and kind show a clear positive correlation between situational analogy and precision. A weaker effect can be observed in the case of the distributive/collective distinction. This is not totally surprising, as the system is designed with the quantification task in mind. The distributive/collective annotation task, despite its relation to quantification resolution, is sufficiently different that it would necessitate its own separate classification module.

It is important to realise that the observed correlation is dependent on the use of syntax. We argued in Section 6.2 that the exact interpretation of a script varies with the syntactic environment in which it occurs. A consequence of this is that there can be no assumed correlation between similarity alone and precision. Figure 6.2 confirms this assumption: it shows the learning curve of the system when running on similarity alone, without any syntactic information. As expected, correlation is weak.

Note that the positive correlation observed for all tasks implies that more significant improvements can be expected when using a larger script database.

The other positive effect of using threshholding is that it is possible to increase the precision of classification for individual classes. We commented at the end of Chapter 5 that one of the big drawbacks of using syntactic features alone was that no further improvement was possible, given the same set of features. Using similarity, however, increasing the amount of data available as we increase the threshold of the system can result in better performance. This effect of threshholding is visible in the precisions calculated for class ONE, where figures rise from 76% at the threshold of 0 to 95% at the threshold of 0.5. Given the small amount of data available, we are unable to show the same effect for the other classes, but we expect that a larger script database would confirm this result.

	No rules	0.1	0.2	0.3	0.4	0.5
Overall precision	66%	71%	73%	75%	77%	77%
Precision of situational analogy	66%	73%	76%	79%	83%	83%
Recall of situational analogy	66%	56%	50%	36%	20%	11%
F-score	66%	63%	60%	49%	32%	19%
ONE	76% (329/434)	88% (289/326)	89% (253/283)	91% (191/209)	93% (105/112)	95% (57/60)
SOME	32% (12/38)	33% (12/36)	29% (8/27)	26% (4/15)	44% (4/9)	33% (2/6)
MOST	38% (26/10)	45% (10/22)	57% (8/14)	85% (6/7)	66% (2/3)	– (0/0)
ALL	38% (33/88)	35% (28/78)	37% (22/59)	34% (12/35)	44% (8/18)	44% (4/9)
QUANT	71% (10/14)	66% (8/12)	66% (8/12)	57% (4/7)	50% (2/4)	0% (0/1)
Precision of rules	–	63%	66%	72%	74%	77%
Recall of rules	–	13%	23%	39%	56%	67%
F-score	–	22%	34%	51%	64%	72%
ONE	–	68% (42/61)	75% (85/112)	78% (156/199)	82% (249/302)	84% (302/358)
SOME	–	58% (7/12)	58% (10/17)	65% (17/26)	56% (21/37)	58% (23/39)
MOST	–	0% (0/3)	0% (0/4)	20% (1/5)	20% (1/5)	33% (2/6)
ALL	–	44% (16/36)	46% (27/58)	53% (42/79)	52% (47/90)	52% (52/99)
QUANT	–	100% (14/14)	100% (14/14)	100% (18/18)	100% (20/20)	100% (22/22)

Table 6.5: The effect of thresholding. Results for the quantification annotation.

	no rules	0.1	0.2	0.3	0.4	0.5
Overall precision	82%	86%	86%	87%	88%	89%
Precision of situational analogy	82%	86%	87%	88%	89%	96%
Recall of situational analogy	82%	67%	57%	40%	22%	12%
F-score	82%	75%	69%	55%	35%	21%
NOTKIND	86% (472/549)	91% (387/424)	91% (327/358)	93% (229/246)	94% (126/133)	97% (71/73)
KIND	27% (10/37)	28% (9/32)	21% (5/23)	30% (6/20)	22% (2/9)	100% (2/2)
QUANT	71% (14/10)	61% (8/13)	66% (8/12)	57% (4/7)	50% (2/4)	0% (0/1)
Precision of rules	–	85%	86%	87%	88%	88%
Recall of rules	–	19%	30%	48%	66%	77%
F-score	–	31%	44%	62%	75%	82%
NOTKIND	–	82% (87/106)	84% (153/182)	85% (256/298)	86% (367/423)	87% (430/491)
KIND	–	–	–	–	–	–
QUANT	–	100% (25/25)	100% (25/25)	100% (29/29)	100% (31/31)	100% (33/33)

Table 6.6: The effect of thresholding. Results for the kind annotation.

	no rules	0.1	0.2	0.3	0.4	0.5
Overall precision	79%	82%	83%	83%	83%	83%
Precision of situational analogy	79%	85%	87%	87%	87%	83%
Recall of situational analogy	79%	67%	57%	40%	22%	11%
F-score	79%	75%	69%	55%	35%	19%
DIST	84% (433/514)	91% (364/396)	92% (310/335)	93% (225/241)	91% (121/132)	90% (63/70)
COLL	40% (29/72)	39% (25/63)	43% (19/44)	33% (9/27)	35% (5/14)	33% (3/9)
QUANT	71% (10/14)	78% (11/14)	78% (11/14)	71% (5/7)	75% (3/4)	0% (0/1)
Precision of rules	–	72%	76%	79%	82%	83%
Recall of rules	–	15%	26%	43%	62%	72%
F-score	–	25%	39%	56%	71%	77%
DIST	–	68% (60/87)	77% (121/156)	81% (209/258)	85% (319/375)	85% (379/441)
COLL	–	59% (13/22)	54% (18/33)	53% (23/43)	54% (27/50)	52% (27/51)
QUANT	–	100% (18/18)	100% (18/18)	100% (24/24)	100% (25/25)	100% (28/28)

Table 6.7: The effect of thresholding. Results for the distributive/collective distinction.

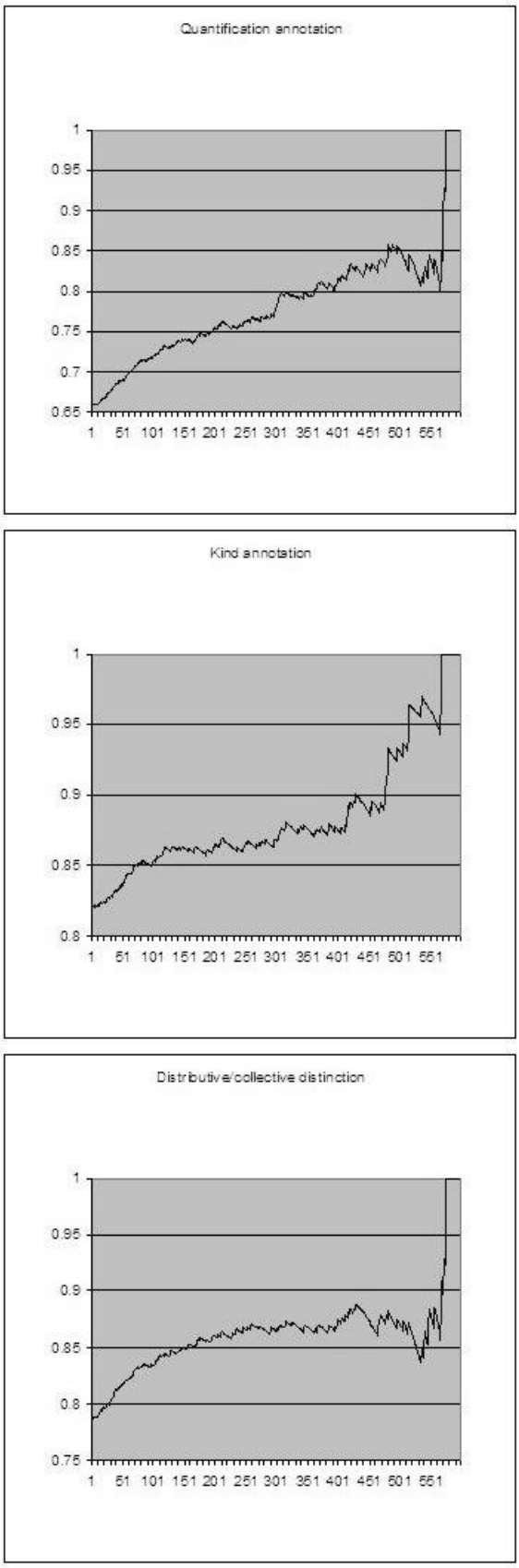

Figure 6.1: Precision against threshold

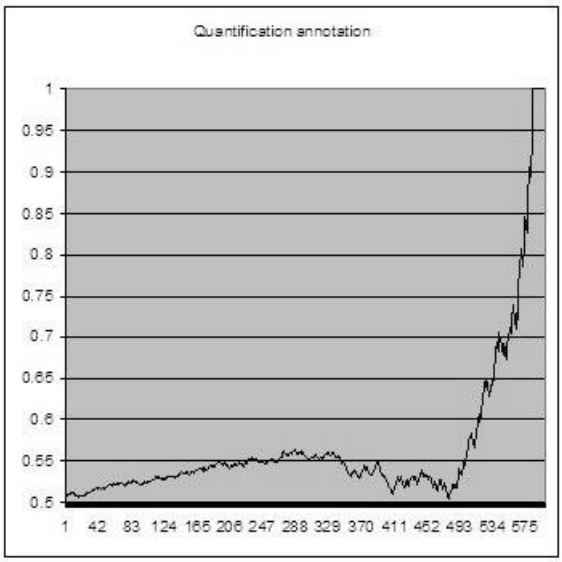

Figure 6.2: Precision against threshold for similarity only

Output 7 Example output for nearest neighbours algorithm, k=3

```
**********************************************************************
Annotating: loyalist leave settlement
**********************************************************************
Scores by class (weighted class score, number of triples):

one: 0 0
some: 0.555703 2
most: 0 0
all: 0.325735 1
quantified: 0 0

**********************************************************************
loyalist leave settlement SCORE 0.63045 ANNOT some
**********************************************************************
```

6.5 Issues with nearest neighbour algorithm

The nearest neighbour algorithm presents an obvious problem: it suffers from noise (the most similar script may have been incorrectly annotated or it may simply be an outlier). In order to counteract the effect of noise, it is possible to move to a k nearest neighbours solution, with majority vote over a group of similar scripts. The problem with this solution is to choose the right k. In our setup, where syntax and similarity are interdependent, setting k high enough that it counteracts the noise in the data also implies that the classifier reverts to a majority class algorithm where each new data point is simply labelled

as the class for which its syntactic context is the most frequent. The benefit of similarity is lost, and the classifier is expected to reach a plateau when the data approximates the 'true' distribution of the syntactic contexts.

In what follows, we report the results of experiments for various values of k, with threshold 0 (no rules are used) and full use of syntax. To overcome the problems associated with high ks, we weight the algorithm using the distance of each nearest neighbour to the statement to be classified. We calculate, as before, the similarity between statement S_0 and all reference scripts $S_{1...n}$, to obtain n situational analogy scores, $a_{1...n}$. We then store each score a_k under the relevant annotation label, as obtained from the corresponding reference script S_k. We then sum all scores under each label to obtain weighted class scores. The separate class scores, in turn, are added up to give the total analogy A for S_0. We finally compute an end score for each label, as the percentage of its weighted class score in A. The label with the highest score gives the annotation for S_0. An example is shown in Output 7 for k=3.

Tables 6.8 to 6.10 show that increasing the value of k has a positive effect on the precision of the system, in particular for minority classes (the improvement is particularly drastic in the case of the KIND class, for which precision goes from 27% to 52%). A plateau is quickly reached, though, for k=5. Looking at the output of the program at k= 10 for Fold 1, we note that due to data sparsity, it is actually relatively rare for the system to find 10 scripts matching the syntactic environment of the triple to be classified (this only happens in a third of all cases). Further, half of the instances to be classified return 6 or less matching scripts. The consequence of this is that the classification may be over-influenced by the distribution of the syntax in our small corpus. We concludes that, although the switch to k neighbours is desirable, it is not clear to what extent it would influence the performance of our system given a larger training corpus.

	k=1	k=3	k=5	k=10
Precision	66%	68%	69%	69%
ONE	76% (329/434)	76% (340/445)	76% (343/451)	76% (345/453)
SOME	32% (12/38)	43% (17/40)	43% (15/35)	47% (15/32)
MOST	38% (10/26)	30% (9/30)	32% (9/28)	29% (8/28)
ALL	38% (33/88)	46% (34/74)	46% (35/76)	47% (36/77)
QUANT	71% (10/14)	91% (10/11)	100% (10/10)	100% (10/10)

Table 6.8: The effect of increasing ks on the best-neighbours algorithm. Results for the quantification annotation.

	k=1	k=3	k=5	k=10
Precision	82%	85%	86%	86%
NOTKIND	86% (472/549)	87% (485/558)	87% (488/560)	87% (489/561)
KIND	27% (10/37)	47% (14/30)	50% (15/30)	52% (15/29)
QUANT	71% (10/14)	91% (10/11)	100% (10/10)	100% (10/10)

Table 6.9: The effect of increasing ks on the best-neighbours algorithm. Results for the kind annotation.

	k=1	k=3	k=5	k=10
Precision	79%	81%	81%	81%
DIST	84% (433/514)	84% (444/526)	84% (450/534)	84% (451/535)
COLL	40% (29/72)	46% (29/63)	48% (27/56)	49% (27/55)
QUANT	71% (10/14)	91% (10/11)	100% (10/10)	100% (10/10)

Table 6.10: The effect of increasing ks on the best-neighbours algorithm. Results for the distributive/collective distinction.

6.6 Beyond nearest neighbours?

The k nearest neighbour algorithm is the most direct implementation of memory-based learning. As mentioned previously in Section 6.2, it is a solution that assumes the importance of outliers in achieving good precision. We have seen in Section 6.5, however, that a strict implementation where $k = 1$ is not necessarily the best and higher values of k may actually provide better results, i.e. some generalisation may be useful. In this section, we briefly investigate whether the need for generalisation outweighs the benefits given by considering outliers in the data and develop an implementation of our system based on **support vector machines** (SVMs, Cortes and Vapnik, 1995). SVMs are a family of supervised classifiers that, given some labelled training data, construct a 'soft' hyperplane boundary to separate the data. The hyperplane can be calculated in a higher dimensional space where the data becomes separable. It should have maximal distance to the nearest points of both classes to be separated. In cases of non-separable data, the classifier constructs the hyperplane that minimises the error margin for the data. Figure 6.3 shows an example of a decision boundary found by an SVM.

In mathematical terms, we search for the hyperplane that satisfies

$$c_i(\mathbf{w} \quad \mathbf{x_i} \quad b) \quad 1 \quad \xi \ _i \quad 1 \quad i \quad n \tag{6.5}$$

where c_i is the label (0 or 1) of data point $\mathbf{x_i}$, $\mathbf{w}$ and b define the hyperplane in relation to each data point $\mathbf{x_i}$ (a hyperplane can be expressed as the set of points $\mathbf{x}$ for which the

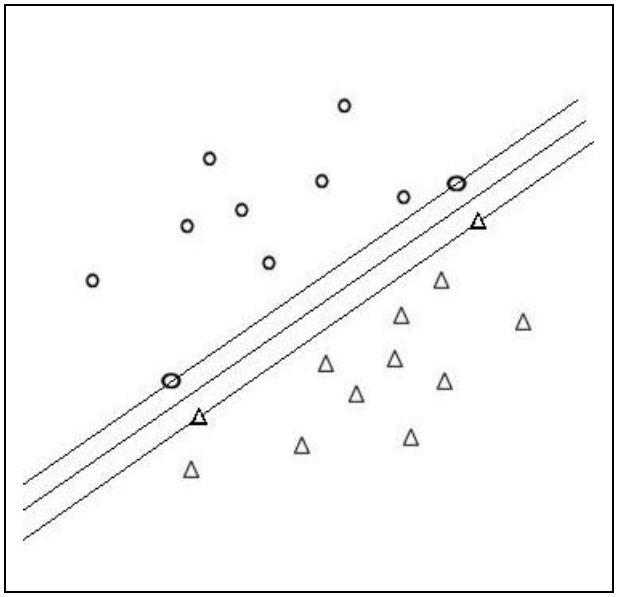

Figure 6.3: Decision boundary found by an SVM with linear kernel

normal vector $\mathbf{w}$ and the parameter b satisfy $\mathbf{w} \quad \mathbf{x} \quad b = 0$) and ξ_i is a slack variable that measures misclassification.

The overall optimisation problem is to find a minimum for

$$\frac{1}{2} \ \mathbf{w} \ ^2 + C \sum_i \xi_i \tag{6.6}$$

where C is a constant expressing the 'cost parameter' or 'error penalty' for the problem.

It is possible to rewrite the term $\mathbf{w}$ ^{2} in 6.6 as a function of the training data points $\mathbf{x}_{i...n}$ involving the dot product function $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \quad \mathbf{x}_j$. This function is called the **kernel** function of the classifier and can be replaced with non-linear functions if desired.

Support vector machines are designed for binary classification, but it is possible to modify them for multiclass problems (Vapnik, 1998). Ó Séaghdha (2008), for instance, uses SVMs with a range of novel distributional kernels to classify noun compounds into semantic relations such as 'location' or 'possession'.

There are constraints on the form that kernels can take. Ó Séaghdha (2008) shows that the Lin measure (1998) — which we used in our similarity calculations as described in Section 6.3.2 — is not a suitable kernel for an SVM. Therefore, in the experiments that follow, we use as input the raw feature vectors calculated for each word in our corpus as per Section 6.3.2. It is easy to show that this representation implicitly forces the SVM to use similarity as basis for classification: our input space has many dimensions which each correspond to a feature in our corpus. Each feature vector is a point in that space. Given this, we can show that the data point for *cat* is closer to *dog* than, say, *bank* (in virtue of shared contexts such as *feed*, *run* or *pet*).

In what follows, we use an off-the-shelf implementation of support vector machines: LIBSVM (Chang and Lin, 2001). As we want to consider similarity between Subject-Verb-Object triples rather than single words, we feed the SVM with the concatenation of the three feature vectors that make up each triple. The data is first normalised in the range [0,1] using the scaling option of LIBSVM, following the recommendation of Hsu et al (2010). The SVM cost parameter C is adjusted by 5-fold cross-validation on the training data, over the range $[2^{-6}, 2^{-4}, ...2^{10}, 2^{12}]$. We use the simple linear kernel option provided by LIBSVM and compute results over the six folds previously used for evaluation.

We run our SVM implementation over the distributional vectors only, and then with syntactic information appended to each vector. The two runs correspond to our 'similarity only' and 'similarity with full syntax' experiments of Section 6.4. Because support vector machines must take numerical input, we convert each syntactic label using n binary attributes for an n-values category, following Hsu et al (2010): the representation of syntactic number, for instance, takes two attributes; the combination $(1, 0)$ indicates a singular while $(0, 1)$ indicates a plural entity.

Our results are shown in Table 6.11, together with figures for a syntax-only experiment given as a baseline for the SVM method. The syntax-only experiment produces identical precision to the tree-based classifier used in Chapter 5. The experiments involving similarity, though, show worse performance than the simple k nearest neighbours algorithm described in Section 6.3.3, which gives 51% precision in the experiments without syntax and up to 77% using full syntax.

	No syntax	Full syntax	Syntax only
Precision	48%	48%	78%
ONE	62% (277/445)	63% (273/435)	86% (359/419)
SOME	0% (0/22)	5% (1/20)	62% (24/39)
MOST	13% (4/32)	10% (4/39)	13% (1/8)
ALL	7% (4/60)	9% (6/69)	55% (56/102)
QUANT	5% (2/41)	5% (2/37)	94% (30/32)

Table 6.11: SVM results for the quantification annotation.

Trying to find a reason for the superiority of the k nearest neighbours method over the support vector machines involves coming back to the representation of our data. We suggested that two triples that share a large part of their contextual distribution are similar, and our nearest neighbour experiments showed that two similar triples in the same syntactic environment tend indeed to have the same quantifier. Although this is true, there is nothing preventing two triples that are a great distance apart in the similarity space from also sharing a quantifier. That is, we can be confident that *Dogs are canines* and *Cats are felines* are similarly quantified because of their proximity but we cannot

say anything about *Quarks are particles* on the ground of its distance to the former two statements (in fact, they all share the same quantifier, *all'*). In other words, quantifier relations are not separable in the semantic space like, for example, the compound relations identified by Ó Séaghdha (2008). In the compound case, the relation itself has a specific semantic distribution, like a content word: we can imagine that the relation SUBSTANCE, which applies to both *glass table* and *steel knife*, is expressed by such features as *... made of...*, *... produced out of strong...* or again *... manufactured from quality....* The same case is not applicable to the SOME relation. We will say that quantifiers do not have a **contiguous** distributional semantics.[1]

This aspect of the nature of quantifiers may explain why a classifier trained to separate data along a (relatively) simple function will fail to do so in a case where each class is scattered across the semantic space. Figure 6.4 shows the difference in complexity between a case of contiguous distribution and one of non-contiguous distribution. In this setup, the k nearest neighbours algorithm may perform better because it only picks out a small area of the whole semantic space for classification. In principle, it should be possible for an SVM to learn a nearest-neighbour-like decision boundary using a more complex kernel function. However, finding the correct choice of parameters for such a problem is nontrivial. In preliminary experiments involving a Gaussian kernel, we found that the classifier was unable to compute a decision boundary for the data and kept classifying all instances in the same category. Further experiments would be needed to set appropriate parameters for the semantic space of quantification.

We should note here that this discussion is speculative and other factors may have contributed to the poor performance of the SVM, in particular the small amount of data available. Further work should involve a more detailed analysis of the results and comparison with other methods, including other k nearest neighbour algorithms (for instance the latest version of TiMBL, the implementation used by Daelemans et al, 1999 — see Section 6.2).

Two conclusions can be drawn from the experiments in this chapter. First, the simple baseline calculated in Chapter 5 is difficult to beat, even when using more complex methods. We believe, however, that a larger annotated corpus would help us gain further precision, in particular when quantifying, and make a significant difference to the results obtained by a system based on syntax alone. Secondly, we can demonstrate that pragmatics plays a role in the resolution of underquantification. Using a problem-solving methods such as analogy indicates a positive correlation between the scores obtained by the system and the quality of the annotation.

It should be clear that our attempts at quantification resolution are only first steps taken in solving a complex problem. There may, for instance, be some gains to be had in the correct processing of rule-like statements where pure lexical semantics can help us quantify

[1] Interestingly, this remark will apply to other function words.

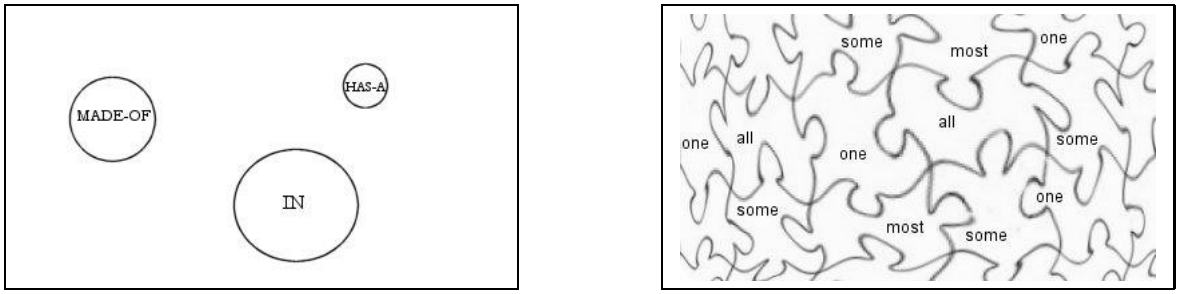

Figure 6.4: Left: the distribution of noun compound relations in the semantic space. Right: the distribution of natural language quantifiers in the semantic space.

outside of context. Furthermore, the right kind of machine learning algorithm must be found to deal with the fact that quantifiers are to some extent 'function' words which resist a direct distributional specification. This, together with any other investigation, will be left for future work.

Chapter 7

Conclusion

We have shown in this thesis that the quantification of noun phrases is an interesting topic, at various levels of investigation. It involves theoretical issues closely related to the study of genericity and thereby brings up the semantic interpretation problems encountered in the analysis of generic noun phrases. The expression of ambiguity, or rather underspecification, necessitates a complex formalisation. On the experimental side, the resolution of quantification leads to the interesting issue of how humans perform such resolution, or sometimes, how they fail in their attempts. Machine learning experiments with different algorithms show that dealing with the semantics of quantifiers involves a view of meaning quite different from that assumed for content words. They also confirm the need for pragmatics in the resolution process.

In this last chapter, we summarise our main contributions with regard to the study of quantification as a linguistic object and as a challenge for Natural Language Processing. We then informally assess how our implementation of a quantification resolution system goes some way in supporting more general NLP tasks such as commonsense statements extraction and inference. We conclude by highlighting the issues left open by our work, which will serve as future research topics.

7.1 Contributions

We have provided in this thesis an investigation of the various aspects of the underquantification phenomenon, from the point of view of computational linguistics. Our claims cover a range of questions, from theory to implementation. We have shown the following:

> All noun phrases can be partially interpreted in terms of quantification. Where quantification is ambiguous, we speak of underquantification, that is the underspecification of the quantifier value in the noun phrase. Genericity phenomena can be analysed in terms of underquantification. So can definite plurals.

It is possible to give a unified formalisation to the underspecified quantifier. Its resolution consists in appending the formalisation with the correct set relation for the observed statement. The fully resolved formalisation involves knowing the distributive or collective status of the verbal predicate and the kind status of the noun phrase.

Humans perform quantification resolution with substantial agreement, although with more noticeable difficulties when having to distinguish between SOME, MOST and ALL labels.

Syntax on its own is insufficient to perform high-quality quantification resolution, although it provides a respectable baseline for the quantification task itself and the classification of distributives against collectives. It is, however, unable to deal with kind annotations.

Quantification resolution is mostly situated at the level of pragmatics. It is to some extent possible to replicate the problem-solving skills used by humans when resolving quantification by implementing a system based on situational analogy. A positive correlation can be observed between the precision of the system on the quantification task and situational analogy.

Separating quantifier classes in a discursive feature space requires more than a simple boundary. This is because quantifiers, like other function words, do not have a contiguous distributional semantics. The consequence of this is that a k nearest neighbours algorithm gives better performance than a linear SVM on the task.

7.2 Quantification resolution in the real world

We claimed in the introduction to this thesis that the quality of large ontologies and databases could be greatly improved by quantification resolution. We have provided a machine-processable formalisation of quantification which could be directly integrated in commonsense extraction systems. Our implementation of the ambiguity resolution process, however, should be seen as no more than a first step in a novel line of research. It will, no doubt, will be followed by many more. Still, we have shown that a situational analogy system, combined with an appropriate threshold, can offer precision and recall at various levels. In all cases where accuracy is the foremost concern, it is possible to provide annotations at over 80% precision and leave the statements not covered by the situational analogy module unannotated.

As for the issue of reasoning over ontologies, it is clear that the formalisation proposed in this thesis allows for inference at instance level: from the representation of *All cats are mammals*, we can easily infer that a particular cat is a mammal (the set relation at the

end of the formalisation can be straightforwardly mapped to probability adverbs as shown in Section 1.1.2). A less trivial problem is performing inference over chains of quantified statements. We will leave the issue as a topic for further research but wish to show that our formalisation lends itself to adequate logical operations: we will demonstrate that if *all trouts are fish* and *all fish live in water*, then *all trouts live in water*. The two sentences are reproduced below, together with their formalisation.

238. All trouts are fish.

$$T = \sigma^* x\, trout'(x) \quad A[A \prod T \quad z[z \quad \prod A \quad fish\ '(z)] \quad T \quad A \ = 0]$$

239. All fish live in water.

$$F = \sigma^* x\, fish'(x) \quad B[B \prod F \quad z[z \quad \prod B \quad liveInWater\ '(z)] \quad F \quad B \ = 0]$$

Let us consider first Sentence 238 and its formalisation. We assume that a supremum F exists, which satisfies $F = \sigma^* x\ fish'(x)$ (that is the same F as in Sentence 239). The fragment of formalisation $z[z \quad \prod A \quad fish\ '(z)]$ then implies $A \prod F$ (if all trouts in A have the property *fish'*, then A is a part of the plurality of all fish F – we assume the axiom '*fish'* is distributive'). Because $A = T$ in virtue of $T \quad A \ = 0$ and $A \prod T$, we can write $T \prod F$.

Sentence 239 can be similarly analysed and, given the supremum of all things that live in water, $W = \sigma^* x\, liveInWater'(x)$, we can write $F \prod W$.

The individual-part operator $\prod$ is transitive, therefore we can deduce: $T \prod F \quad F \prod W \quad T \prod W$ (the supremum of trouts is a part of the supremum of all things that live in water). *liveInWater'* is a distributive property so we conclude that $z[z \quad \prod T \quad liveInWater\ '(z)]$ (all trouts live in water).

There are obvious difficulties in chaining statements that are not all universally quantified, such as *All penguins are birds* and *Most birds fly* (see Bacchus, 1989, for an attempt to provide a statistical interpretation of such sentences in a reasoner). We should note, however, that some state-of-the-art inference systems successfully deal with already qualified statements (MacCartney and Manning, 2008). Good quantification resolution could give dramatic coverage improvements to such systems.

7.3 Remaining issues

Our work leaves several implemention issues open. First, we haven't been able to show substantial improvements in precision over our syntactic baseline when introducing more complex semantic methods. A correlation does exist between performance and situational analogy score but it is not clear how much more training data would be needed to actually

achieve increases of several percents in the results. The first avenue for future work would therefore be the organisation of a fairly large annotation effort with a view to confirm the usefulness of situational analogy. We hope that scaling our script database from 500 instances to several thousands would produce a dramatic increase in performance. From the point of view of close investigation, a larger database would also allow us to answer questions that were left open by our experiments. In particular, we could assess the benefit of increasing k in our use of the k nearest neighbours algorithm (see Section 6.5). We could also verify that putting a threshold on analogy scores increases individual class precisions (Section 6.4). Overall, further work is needed to identify a truly adequate classification algorithm for the task and additional data would be welcome for our investigation.

Secondly, we remarked that the system was comparatively weak in labelling the SOME, MOST and ALL classes. We can attribute the problem to data sparsity, but it is also potentially a consequence of the annotation being less reliable on those classes. A new annotation should involve a revised scheme that addresses the problems highlighted in Section 4.5.3, by helping annotators resolve the noun phrase's referent, encouraging the use of external resources and spelling out rules to deal with temporal quantification. Issues linked to reference should be treated with particular caution, as the design of guidelines in this respect will affect the general usability of the annotation. Specifically, we can say that there is a correlation between the extent to which humans perform reference resolution and the assumed complexity of the systems which will be trained on the produced data. The following example should make this clear:

240. The whale is a mammal. [...] *The animal* feeds on plankton.

Taking the lexical realisation of the noun phrase as reference produces the annotation SOME: *Some animals feed on plankton.* Asking annotators to resolve the anaphora would result in the noun phrase being labelled as ALL: *All whales feed on plankton.* The latter assumes that whichever system is trained on this example would include an anaphora resolution module. While arguments can be made in favour of more complex resolution (because it doesn't result in information loss), it may be sensible to first focus on usability — or to produce multiple annotations of various complexity.

Finally, we explained in our introduction why we left object noun phrases out of our investigation. In order to give a truly compositional formalisation of main clauses in text, we must attempt to find a consistent account for the quantification of objects. As we have already commented, the issue is debated as a very fundamental level in the linguistic literature. It may however be possible to build a working computational model for those constructs, and this should be one of our next theoretical goals.

Bibliography

[1] ACE. 2008. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*, Version 6.6 2008.06.13. Linguistic Data Consortium.

[2] Nicholas Asher and Michael Morreau. 1991. 'Commonsense entailment: A modal theory of non-monotonic reasoning'. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, San Mateo, CA, pages 387–392. San Francisco, CA: Morgan Kaufmann.

[3] Fahiem Bacchus. 1989. 'A modest, but semantically well founded, inheritance reasoner'. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Detroit, MI, pages 1104–1109. San Francisco, CA: Morgan Kaufmann.

[4] Leila Behrens. 2005. 'Genericity from a Cross-linguistic Perspective'. *Linguistics*, 43(2):275–344.

[5] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka. 2007. 'Measuring semantic similarity between words using web search engines'. In *Proceedings of the 16th international conference on the World Wide Web*, Banff, Alberta, Canada, pages 757–766.

[6] Edward Briscoe, John Carroll and Rebecca Watson. 2006. 'The Second Release of the RASP System'. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistic(COLING/ACL-06), Interactive Presentation Sessions*, Sydney, Australia.

[7] Berit Brogaard. 2007. 'The But Not All: A Partitive Account of Plural Definite Descriptions'. *Mind and Language*, 22:402–426.

[8] Noel Burton-Roberts. 1977. 'Generic sentences and analyticity'. *Studies in Language*, 1:155–196.

[9] Gregory Carlson. 1977. *Reference to Kinds in English*. Ph.D. Dissertation. University of Massachusetts at Amherst. Published by Garland, New York, 1980.

[10] Gregory Carlson. 1995. 'Truth-conditions of generic sentences: Two contrasting views'. In Gregory N. Carlson and Francis Jeffry Pelletier, Editors, *The Generic Book*, pages 224–237. Chicago: Chicago University Press.

[11] Helen M. Cartwright. 1975. 'Some remarks about mass nouns and plurality'. In Francis J. Pelletier, Editor, *Mass Terms*, pages 31–46. Dordrecht, Holland: D. Reidel Publishing Company.

[12] Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: A library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] Gennaro Chierchia. 1998. 'Reference to kinds across languages'. *Natural Language Semantics*, 6:339–405.

[14] Timothy Chklovski and Patrick Pantel. 2004. 'VerbOcean: Mining The Web for Fine-Grained Semantic Verb Relations'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain, pages 33–40.

[15] Andrei Cimpian and Ellen M. Markman. 2008. 'Preschool children's use of cues to generic meaning'. *Cognition*, 107(1):19–53.

[16] Ariel Cohen. 1996. *Think Generic: The Meaning and Use of Generic Sentences*. Ph.D. Dissertation. Carnegie-Mellon University at Pittsburgh. Published by CSLI Publications, Stanford, 1999.

[17] Ariel Cohen. 2001. 'Relative Readings of Many, Often, and Generics'. *Natural Language Semantics*, 9(1):41–67.

[18] Ariel Cohen. 2002. 'Genericity'. *Linguistische Berichte*, 10:59–89.

[19] Ariel Cohen and Nomi Erteschik-Shir. 2002. 'Topic, Focus, and the Interpretation of Bare Plurals'. *Natural Language Semantics*, 10(2):125–165.

[20] Jacob Cohen. 1960. 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement*, 20(1):37–46.

[21] Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier and Karsten Konrad. 1996. *Using the Framework*. The Fracas Consortium.

[22] Ann Copestake. 1989. 'Some Notes on Mass Terms and Plurals'. Technical Report 190. University of Cambridge, Computer Laboratory.

[23] Ann Copestake. 2004. 'Robust Minimal Recursion Semantics'. Available at `www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf`. Last accessed 6th April 2010.

[24] Ann Copestake and Ted Briscoe. 1995. 'Semi-productive Polysemy and Sense Extension'. *Journal of Semantics*, 12:15–67.

[25] Ann Copestake, Dan Flickinger, Ivan Sag and Carl Pollard. 2005. 'Minimal Recursion Semantics: An introduction'. *Journal of Research on Language and Computation*, 3(2-3):281–332.

[26] Corinna Cortes and Vladimir Vapnik. 1995. 'Support vector networks'. *Machine Learning*, 20(3):273-297.

[27] William Croft and D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge, UK: Cambridge University Press.

[28] Walter Daelemans, Antal Van Den Bosch and Jakub Zavrel. 1999. 'Forgetting Exceptions Is Harmful in Language Learning'. *Machine Learning*, 34(1-3):11–41. Hingham, MA: Kluwer Academic Publishers.

[29] Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein and Carlo Strapparava. 2006. 'Direct Word Sense Matching for Lexical Substitution'. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistic (COLING/ACL-06)*, Sydney, Australia, pages 449–456.

[30] Östen Dahl. 1975. 'On Generics'. In Edward L. Keenan, Editor, *Formal Semantics of Natural Language*, pages 99–111. Cambridge, UK: Cambridge University Press.

[31] Svetoslav Dankov, Rafal Rzepka and Kenji Araki. 2008. 'Commonsense and context: a novel approach for automatic extraction of generic statements'. In *Proceedings of The 22nd Annual Conference of the Japanese Society for Artificial Intelligence (JSAI-08)*, Asahikawa, Japan, CD-ROM Proceedings: 2P2–6.

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

[33] David Dowty. 1987. 'Collective predicates, distributive predicates and *all*'. In Fred Marshall, Ann Miller and Zheng-sheng Zhang, Editors, *Proceedings of the Third Eastern States Conference on Linguistics*, pages 97–115. Columbus: The Ohio State University, Department of Linguistics.

[34] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld and Alexander Yates. 2004. 'Web-Scale Information Extraction in KnowItAll'. In *Proceedings of the 13th international conference on World Wide Web*, New York, NY, pages 100–110.

[35] Alvan R. Feinstein and Domenic V. Cicchetti. 1990. 'High agreement but low kappa: I. The problems of two paradoxes'. *Journal of Clinical Epidemiology*, 43(6):543–549.

[36] Fleiss. 1971. 'Measuring nominal scale agreement among many raters'. *Psychological Bulletin*, 76(5):378-382.

[37] Maayan Geffet and Ido Dagan. 2005. 'The Distributional Inclusion Hypothesis and Lexical Entailment'. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 107–114.

[38] Susan A. Gelman. 2004. 'Learning words for kinds: Generic noun phrases in acquisition'. In D. Geoffrey Hall and Sandra R. Waxman, Editors, *Weaving a lexicon*. Cambridge, MA: MIT Press.

[39] Claudia Gerstner-Link and Manfred Krifka. 1993. 'Genericity'. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld and Theo Vennemann, Editors, *Syntax: An International Handbook of Contemporary Research*, pages 966–978. Berlin: de Gruyter.

[40] Claudio Giuliano and Alfio Gliozzo. 2007. 'Instance Based Lexical Entailment for Ontology Population'. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pages 248–256.

[41] Sheila R. Glasbey. 2007. 'Bare plurals in object position: which verbs fail to give existential readings, and why?' In Liliane Tasmowski and Svetlana Vogeleer, Editors, *Non-definiteness and Plurality*. John Benjamins Publishing Company, Linguistics Today series.

[42] H. Paul Grice. 1975. 'Logic and Conversation'. In Peter Cole and J. L. Morgan, Editors, *Syntax and Semantics*, Volume 3. New York: Academic Press.

[43] Zelig Harris. 1954. 'Distributional Structure'. In *Word*, 10(2–3):146–162.

[44] Marti Hearst. 1992. 'Automatic Acquisition of Hyponyms from Large Text Corpora'. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING92)*, Nantes, France, pages 539-545.

[45] Aurelie Herbelot. 2009. 'Finding Word Substitutions Using a Distributional Similarity Baseline and Immediate Context Overlap'. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Student Session*, Athens, Greece.

[46] Aurelie Herbelot and Ann Copestake. 2006. 'Acquiring Ontological Relationships from Wikipedia using RMRS'. In *Proceedings of the International Semantic Web Conference (ISWC-06), Workshop on Web content Mining with Human Language Technologies*, Athens, GA.

[47] Aurelie Herbelot and Ann Copestake. 2008. 'Annotating Genericity: How Do Humans Decide? (A Case Study in Ontology Extraction)'. In Sam Featherston and Susanne Winkler, Editors, *The Fruits of Empirical Linguistics. Volume 1: Process*, pages 103–122. Berlin: de Gruyter.

[48] Gerhard Heyer. 1990. 'Semantics and Knowledge Representation in the Analysis of Generic Descriptions'. *Journal of Semantics*, 7(1):93–110.

[49] Graeme Hirst and David St-Onge. 1998. 'Lexical Chains As Representations of Context for the Detection and Correction of Malapropisms'. In Christiane Fellbaum, Editor, *WordNet*. Cambridge, MA: The MIT Press.

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

[51] Michelle A. Hollander, Susan A. Gelman and Jon Star. 2002. 'Children's Interpretation of Generic Noun Phrases'. *Developmental Psychology*, 36:883-894.

[52] John Horty, Richmond Thomason and David Touretzky. 1990. 'A Skeptical Theory of Inheritance in Nonmonotonic Semantic Networks'. *Artificial Intelligence*, 42(2-3):311–348.

[53] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2010. *A practical guide to support vector classification*. Technical report, Dept. of Computer Science, National Taiwan University. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. Last accessed 8th April 2010.

[54] Roderick A. Jacobs. 1969. 'Linguistic Universals and Their Relevance to TESOL'. *TESOL Quarterly*, 3(2):117–122.

[55] Mario Jarmasz and Stan Szpakowicz. 2003. 'Roget's Thesaurus and Semantic Similarity'. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 212–219.

[56] Stig W. Jørgensen. 2000. *Computational Reference. An investigation, Development and Implementation of Kronfeld's Theory of Reference.* Ph.D. Dissertation. Copenhagen Business School.

[57] Hans Kamp. 1981. 'A Theory of Truth and Semantic Representation'. In Jeroen Groenendijk, Theo Janssen and Martin Stokhof, Editors, *Formal Methods in the Study of Language*. Amsterdam: Mathematics Center.

[58] Sangeet Khemlani, Sarah-Jane Leslie, Sam Glucksberg and Paula Rubio-Fernandez. 2007. 'Do ducks lay eggs? How people interpret generic assertions'. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 395–401.

[59] Sangeet Khemlani, Sarah-Jane Leslie and Sam Glucksberg. 2008. 'Syllogistic reasoning with generic premises: The generic overgeneralization effect'. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Washington DC, pages 619–624.

[60] Paul Kingsbury, Martha Palmer and Mitch Marcus. 2002. 'Adding Semantic Annotation to the Penn TreeBank'. In *Proceedings of the Human Language Technology Conference (HLT-02)*, San Diego, CA, pages 252–256.

[61] Manfred Krifka. 1987. *An outline of genericity*. Technical Report SNS-Bericht 87-25. Seminar für natürlich-sprachliche Systeme, Tübingen University, Germany.

[62] Manfred Krifka. 1995. 'Focus and the interpretation of generic sentences'. In Gregory N. Carlson and Francis Jeffry Pelletier, Editors, *The Generic Book*, pages 238–264. Chicago: Chicago University Press.

[63] Manfred Krifka. 2004. 'Bare NPs: kind-referring, indefinites, both, or neither?' In Olivier Bonami and Patricia Cabredo Hofherr, Editors, *Empirical Issues in Formal Syntax and Semantics*, Volume 5, pages 111-132.

[64] Manfred Krifka, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link and Gennaro Chierchia. 1995. 'Genericity: An Introduction'. In Gregory N. Carlson and Francis Jeffry Pelletier, Editors, *The Generic Book*, pages 1–125. Chicago: Chicago University Press.

[65] Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

[66] Benjamin Kuipers. 1979. 'On Representing Commonsense Knowledge'. In Nicholas V. Findler, Editor, *Associative Networks: The Representation and Use of Knowledge by Computers*, pages 393–408. NY: Academic Press.

[67] J. Richard Landis and Gary G. Koch. 1977. 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics*, 33:159–174.

[68] Fred Landman. 1989. 'Groups, I'. *Linguistics and Philosophy*, 12(5):559–605.

[69] Shalom Lappin. 2000. 'An Intensional Parametric Semantics For Vague Quantifiers'. *Linguistics and Philosophy*, 23(6):599–620.

[70] Douglas Lenat. 1990. 'Cyc: Towards Programs with Common Sense'. *Communications of the Association for Computing Machinery*, 33(8):30-49.

[71] Ernest Lepore. 1983. 'What model theoretic semantics cannot do?' *Synthese*, 4(2):167–187.

[72] Sarah-Jane Leslie. 2007. 'Generics and the Structure of the Mind'. *Philosophical Perspectives*, 21(1):375–403.

[73] Sarah-Jane Leslie. 2008. 'Generics: Cognition and Acquisition'. *Philosophical Review*, 117(1):1–47.

[74] Sarah-Jane Leslie, Sangeet Khemlani, Sandeep Prasada and Sam Glucksberg. 2009. 'Conceptual and linguistic distinctions between singular and plural generics'. In Niels Taatgen and Hedderick van Rijn, Editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Holland.

[75] Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

[76] Dekang Lin. 1998. 'An Information-Theoretic Definition of Similarity'. In *Proceedings of the International Conference on Machine Learning (ACM-98)*, Madison, WI, pages 296–304.

[77] Dekang Lin and Patrick Pantel. 2001. 'DIRT – Discovery of Inference Rules from Text. In *Proceedings of the International Conference on Machine Learning (ACM-01)*, San Francisco, CA, pages 323–328.

[78] Godehard Link. 1983. 'The Logical Analysis of Plurals and Mass Terms: a lattice–theoretical approach'. In Rainer Bauerle, Christoph Schwarze and Arnim von Stechow, Editors, *Meaning, Use, and Interpretation of Language*, pages 302–323. Berlin: de Gruyter.

[79] Godehard Link. 1995. 'Generic Information and Dependent Generics'. In Gregory N. Carlson and Francis Jeffry Pelletier, Editors, *The Generic Book*, pages 358–382. Chicago: Chicago University Press.

[80] Godehard Link. 1998. 'Plural'. In *Algebraic Semantics in Language and Philosophy*. Stanford: CSLI Publications.

[81] Christopher Lyons. 1999. *Definiteness*. Cambridge, UK: Cambridge University Press.

[82] Bill MacCartney and Christopher D. Manning. 2008. 'Modeling semantic containment and exclusion in natural language inference'. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK, pages 521–528.

[83] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. 'Building a large annotated corpus of English: The Penn Treebank'. *Computational Linguistics*, 19(2):313–330.

[84] John McCarthy. 1959. 'Programs with common sense'. In *Proceedings of the Teddington Conference on the Mechanisation of Thought Processes*.

[85] James McCawley. 1971. 'Interpretative semantics meets Frankenstein'. *Foundations of Language*, 7:285–296.

[86] Drew McDermott and Jon Doyle. 1982. 'Non-monotonic Logic I'. *Artificial Intelligence*, 13:41–72.

[87] George Miller and Walter Charles. 1991. 'Contextual Correlates of Semantic Similarity'. In *Language and Cognitive Processes*, 6(1):1–28.

[88] Diarmuid Ó Séaghdha. 2008. *Learning compound nouns semantics*. Ph.D. Dissertation. University of Cambridge, United Kingdom.

[89] Patrick Pantel and Marco Pennacchiotti. 2006. 'Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations'. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistic (COLING/ACL-06)*, Sydney, Australia, pages 113–120.

[90] Francis Jeffry Pelletier and Nicolas Asher. 1997. 'Generics and defaults'. In Johan van Benthem and Alice ter Meulen, Editors, *Handbook of Logic and Language*, pages 1125–1177. Amsterdam: Elsevier.

[91] Stanley Peters and Dag Westerståhl. 2006. *Quantifiers in Language and Logic*. Oxford, UK: Oxford University Press.

[92] Massimo Poesio. 2000. 'The GNOME annotation scheme manual', Fourth Version. `http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm`

[93] Massimo Poesio. 2004. 'Discourse Annotation and Semantic Annotation in the GNOME Corpus'. In *Proceedings of the the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Workshop on Discourse Annotation*, Barcelona, Spain.

[94] Sandeep Prasada and Elaine M. Dillingham. 2006. 'Principled and statistical connections in common sense conception'. *Cognition*, 99(1), pages 73–112.

[95] Willard Van Orman Quine. 1960. *Word and Object*. The MIT Press.

[96] John Ross Quinlan. 1993. *Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann.

[97] Deepak Ravichandran and Eduard Hovy. 2002. 'Learning Surface Text Patterns for a Question Answering system.' In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA, pages 41–47.

[98] Dennis Reidsma and Jean Carletta. 2008. 'Reliability measurement without limits'. *Computational Linguistics*, 34(3), pages 319–326.

[99] Ray Reiter. 1980. 'A logic for default reasoning'. *Artificial Intelligence*, 13:81–137.

[100] Philip Resnik. 1995. 'Using Information Content to Evaluate Semantic Similarity in a Taxonomy'. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada, pages 448–453.

[101] Anna Ritchie. 2004. 'Compatible RMRS Representations from RASP and the ERG'. Technical Report UCAM-CL-TR-661. University of Cambridge, Computer Laboratory.

[102] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. 'Multiword Expressions: A Pain in the Neck for NLP'. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02)*, Mexico City, Mexico, pages 1–15.

[103] Roger C. Schank, and Robert P. Abelson. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.

[104] Lenhart K. Schubert and Francis Jeffry Pelletier. 1987. 'Problems in the representation of the logical form of generics, plurals, and mass nouns'. In Ernest LePore, Editor, *New Directions in Semantics*, pages 385–451. London: Academic Press.

[105] Farzad Sharifian and Ahmad R. Lotfi. 2003. 'Rices and waters: The mass/count distinction in Modern Persian'. *Anthropological Linguistics*, 45(2):226–244.

[106] Push Singh. 2002. 'The public acquisition of commonsense knowledge'. In *Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA.

[107] Craig Stanfill and David Waltz. 1986. 'Toward memory-based reasoning'. In *Communications of the Association for Computing Machinery*, 29(12):1213–1228.

[108] Idan Szpektor, Eyal Shnarch and Ido Dagan. 2007. 'Instance-Based Evaluation of Entailment Rule Acquisition'. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Czech Republic, pages 456–463.

[109] Idan Szpektor, Hristo Tanev, Ido Dagan and Bonaventura Coppola. 2004. 'Scaling Web-based Acquisition of Entailment Relations'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain, pages 41–48.

[110] Sangweon Suh. 2006. *Extracting Generic Statements for the Semantic Web*. MSc Dissertation. University of Edinburgh, United Kingdom.

[111] Sangweon Suh, Harry Halpin and Ewan Klein. 2006. 'Extracting Common Sense Knowledge from Wikipedia'. In *Proceedings of the the International Semantic Web Conference (ISWC-06), Workshop on Web Content Mining with Human Language Technology*, Athens, GA.

[112] Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Summarisation and Citation Indexing*. Stanford: CSLI Publications.

[113] Simone Teufel, Advaith Siddharthan and Dan Tidhar. 2006. 'An annotation scheme for citation function'. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia, pages 80–87.

[114] Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. New York: Wiley.

[115] Carl Vogel. 1995. *Inheritance Reasoning: Psychological Plausibility, Proof Theory and Semantics*. Ph.D. Dissertation. University of Edinburgh, United Kingdom.

[116] Carl Vogel. 2008. 'Metaphor is generic'. In Khurshid Ahmad, Editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation, Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, Marrakesh, Morocco.

[117] Carl Vogel and Michelle McGillion. 2002. 'Genericity is conceptual, not semantic'. In *Proceedings of the Seventh Symposium on Logic and Language*, Pecs, Hungary.

[118] Johanna Völker, Pascal Hitzler and Philipp Cimiano. 2007. 'Acquisition of OWL DL Axioms from Lexical Resources'. In *Proceedings of the Fourth European conference on The Semantic Web: Research and Applications*, pages 670–685. Springer-Verlag.

Appendix A

Guidelines for the quantification annotation task

A.1 Material

You will be given a list of 'triples' of the type 'A does B' and the sentence from which each triple was extracted. For instance:

> TRIPLE: state win case
>
> ORIGINAL: Few roads even entered the area before 1929, when the state won another court case and built what is now known as the Pacific Coast Highway.

Your task is to annotate the noun phrase (NP) that contains the first element of the triple, that is, the **subject** of the triple. In the example above, you should annotate 'the state'. Your annotation should follow the following format:

> Triple Number – subject NP – annotation for subject NP

So, for the above example:

> 1 – the state – one

A.2 The task

The annotation that you give for each sentence will depend on instructions given later in these guidelines. Those instructions are given in the form of decision trees and ask you to provide three types of annotations: you are requested to **quantify** the noun phrase under consideration, to tell whether it is a **kind** or not, and to decide whether the verb phrase in the sentence applies to the subject noun phrase **distributively** or **collectively**. To help you, definitions of the terms in bold are provided below.

A.2.1 Some definitions

Quantification

By 'quantifying' we mean paraphrasing the NP using a term expressing some quantity, such as *some* or *most*:

> *Europeans* discovered the Tuggerah Lakes in 1796 = *Some Europeans* discovered the Tuggerah Lakes in 1796.

Note that the meaning of the whole sentence *mustn't* change.

Collective vs Distributive

A distributive statement is one where every entity referred to by the subject is individually involved in the verb's action:

> *The students* took an exam = Each student, individually, took the exam.

A collective statement is one where the group referred to by the subject, as opposed to individuals, performs the action:

> *The residents* founded a self-help group = The residents, together, founded a self-help group (and not: each residents founded their separate self-help group).

Kinds

You will sometimes be asked to annotate a noun phrase as a kind. No decision is needed there, but for your guidance, we define a 'kind' as the entire group of entities described by the noun phrase under consideration:

> *Ducks* lay eggs = The group of all ducks, collectively, lays eggs.

A.3 More on quantification

You will be asked to explicitely quantify each noun phrase. The quantifier should be one of several suggested by the scheme (see Section A.3.1) – try each one and add the most fitting one to your annotation (the quantifier must be added to the noun phrase without changing the meaning of the sentence.) If there is already an article in front of the noun phrase, try using *some of, etc* instead of *some, etc.*

1. [Some] Europeans discovered the Tuggerah Lakes in 1796.
2. [Most of] The organizers of the exhibition were appaled.
3. [All] Such drifts hurt portfolios that are built with diversification as a high priority.

A.3.1 The annotation labels

You will be using five labels in the course of the annotation: *quant, one, some, most* and *all*.

quant: used when a noun phrase is already quantified:

> some people/6 million inhabitants/most of the workers

one: used when the noun phrase refers to an individual, distinguishable entity – or when the entity is unique in the world:

> My cat = a given individual, distinguishable from other cats, possibly also from other cats of mine. I can point at it.
>
> Bohr's model of the atom = a unique entity (there are not several such models).
>
> The Eiffel Tower = a unique entity.

some: self-explanatory. Used when part of a group is involved in the action:

> *Europeans* discovered the Tuggerah Lakes in 1796 = a few individuals taken out of the group of all Europeans.

most: self-explanatory. Used when the majority of a group is involved in the action:

> *Cats* have four legs = most of them – some of them are injured or have a birth defect.

all: self-explanatory. Used when all members of a group are involved:

> *Cats* are mammals = every cat is necessarily a mammal.

Note that cases where the quantifier *both* would normally be used map onto *all*:

> *The brothers* [Romulus and Remus] were good warriors = all of (the two of) them.

A.3.2 What to do when you hesitate?

In general, try to imagine the referent set of the noun phrase – that is, what the NP refers to in the whole sentence, as opposed to what it refers to at the 'point of hearing':

> Recent research papers have described modern dolomite formations under anaerobic conditions.
>
> [When I hear of *recent research papers*, I think of all recent research papers. The sentence is obviously only referring to some of those, so the annotation is *some*.]
>
> The filmmakers achieved the reinstatement of the President, and they founded the Film Directors' Society that same year.
>
> [When I hear of *the filmmakers*, I think of a certain group of filmmakers. The sentence is referring to the whole group, so I annotate as *all*.]
>
> The supermarket round the corner has an offer: pineapples are on three for two.
>
> [By the time I hear of *pineapples*, I think of all pineapples at the local supermarket. The sentence is referring to them as a whole, so I annotate as *all*.]

If you are still hesitating: the label *most* has priority over *all*, and *some* has priority over the other two.

A.4 The annotation process

A different annotation process should be followed depending on the type of noun phrase that you are annotating. There are five basic types: already quantified noun phrases, proper nouns, (non-bare) singulars, plurals and bare singulars. Instructions for each of those types are given below. Simply follow the instructions until you get to the keyword 'finish'. It is helpful to refer to Section A.2.1 to decide between collective and distributive annotations, and to Section A.3.1 for a description of the annotation labels.

A.4.1 Quantified NPs

If the noun phrase is already quantified: annotate as *quant*.

> *Some governments* have labelled the church as a cult.
>
> *Many players* today use plastic plectra.

A.4.2 Proper nouns

For the purpose of this annotation, we define a proper noun as a noun phrase that contains capitalised words and that refers to a concept that doesn't have instances:

1. Proper: *John Smith, Easter Island, World War II...* (Nothing 'is a' John Smith or 'an instance of' World War II)
2. Non proper: *The Romans, The Chicago Bulls* (Caesar is a Roman, Michael Jordan is a Chicago Bull)
3. A case to consider:
   (a) *The First Circle* is an important book = one of Soljenitsyn's novels, no instances possible. Proper.
   (b) He still hasn't returned my *First Circle* = a copy of the book The First Circle. All such copies are instances of the concept 'physical copy of the First Circle'. Non proper.

All proper nouns should be annotated as *one*.

A.4.3 (Non-bare) singulars

Those are singular noun phrases introduced by an article, e.g. *a car*, *the fish and chips shop in the town centre*...

1. Try to pluralise the noun phrase. If it is possible to pluralise the noun phrase and keep the original sentence meaning, go to 3, otherwise go to 2.

   NB: if the noun phrase is an indefinite singular, just bare pluralise it. If the noun phrase is a singular introduced by a demonstrative or possessive article, try to pluralise keeping the article. If the noun phrase is a definite singular, use a bare plural or keep the article, whichever feels most natural.

2. Annotate as *one*. Finish.

   > *The film* featured two songs by Radiohead , Fake Plastic Trees and My Iron Lung taken from the album The Bends. [one]

3. Annotate as *distributive* or *collective*. Go to 4.
4. Annotate as *kind-*. Go to 5.

5. Annotate as *some*, *most* or *all*, with reference to the pluralised sentence. Finish.

   NB: Don't be fooled by sentences with modals:

   > *Individual provinces* may accord their primate with more or less authority. [all] (Don't be biased by the fact that only some or most will indeed do it. All of them **may** do it.)

A.4.4 Plurals

Those are all plural noun phrases, bare or introduced by an article, e.g. *conscripts*, *the founders of the club*...

1. Annotate as either *distributive* or *collective*.

2. Try to singularise the sentence with either *a* or *the*, keeping its meaning. If it is possible, go to 3. Otherwise, annotate as *some*, *most*, or *all* and finish.

   > *Europeans* discovered the Tuggerah Lakes in 1796 = *A European* discovered the Tuggerah Lakes in 1796. [collective/some]

3. Annotate as *kind*.

   > *Voluntary female conscripts* receive a small additional benefit = *A voluntary female conscript* receives a small additional benefit. [distributive/kind/]

4. Annotate as *some*, *most*, or *all*. Finish.

   > *New York residents* founded the theatre in 1928. [collective/some]
   >
   > *Community members* debate all issues once a week. [collective/all]

A.4.5 Bare singulars

Those are singular noun phrases without article, e.g. *psychology*, *water*, *modern Finnish popular music*...

NB: If the bare singular is the result of an ellipsis, add the appropriate article and annotate with reference to the section on non-bare singulars:

> Once deployed, EPIRBs can be activated, depending on the circumstances, either manually (*crewman* flicks a switch) or automatically (as soon as water comes into contact with the unit's see-switch) = a crewman...

Bare singulars are notoriously difficult to annotate. So instead of annotating the actual noun phrase, try first to paraphrase the sentence with a plural noun phrase of your choice (the paraphrase does not have to be perfect, and if you cannot find a synonym of the noun phrase under consideration, use something that is an instance of it):

> *Free software* allows users to co-operate in enhancing and refining the programs they use = *Open source programs* allow users to co-operate...
>
> *Damage* showed seismic resistance deficiencies in modern apartment construction = *Cracks* showed seismic resistance deficiencies...

If you are able to do such a conversion, simply annotate the noun phrase using the section on plurals and write which word(s) you chose as a paraphrase:

> 32 – free software – distributive/kind/all (open source programs)

If the paraphrase is impossible, annotate as *one*:

> Although *modern analytic celestial mechanics* started 400 years ago with Isaac Newton, prior studies addressing the problem of planetary positions are known going back perhaps 3000 or more years. [one]

A.5 Decision trees

This section is here to help you make a decision for each type of noun phrase encountered. Simply follow up the relevant decision tree, referring to the detailed instructions when needed. Decision points enclosed in square boxes indicate that an annotation is needed.

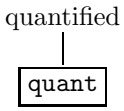

Figure A.1: Quantified case.

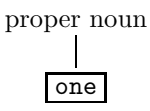

Figure A.2: Proper noun case.

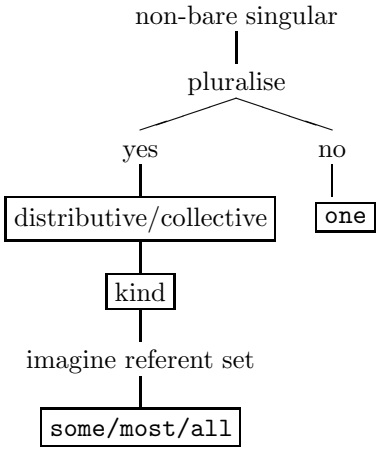

Figure A.3: Non-bare singular case.

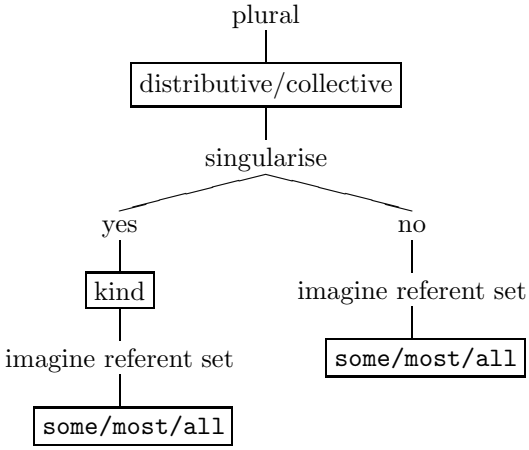

Figure A.4: Plural case.

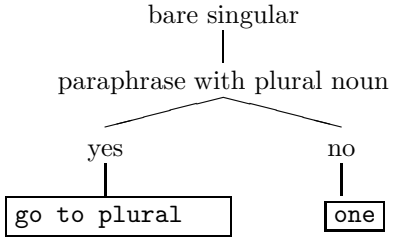

Figure A.5: Bare singular case.

Appendix B

Subset of annotated data

This appendix contains a subset of the annotated data discussed in Chapter 4. It contains the first 50 annotation instances, out of a full set of 300, followed by the decisions made by all three annotators.

Annotation instance	**A1**	**A2**	**A3**
1 ***** TRIPLE: conscript receive training ORIGINAL: The conscripts first receive basic training, after which they are assigned to various units for special training.	ONE KIND DIST	ONE KIND DIST	ONE KIND DIST
2 ***** TRIPLE: pool approach coast ORIGINAL: Warm water pool approaches South American coast.	ONE NOTKIND DIST	ONE NOTKIND COLL	ONE NOTKIND DIST
3 ***** TRIPLE: wing support unit ORIGINAL: The Wing also supports 113 units stretching from Thunder Bay, to the Saskatchewan-Alberta border and from the 49th Parallel to the high Arctic.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
4 ***** TRIPLE: storm turn wave ORIGINAL: Powerful winter storms in the Pacific Ocean can turn typically placid and rolling South Bay waves into large and occasionally dangerous monsters, a natural draw for the local surfing population.	ALL KIND DIST	ALL KIND DIST	ALL KIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
5 ***** TRIPLE: council cover county ORIGINAL: The Pee Dee Area Council covers 11 counties in northeastern South Carolina (Pee Dee): Darlington, Chesterfield, Marlboro, Florence, Dillon, Marion, Horry, Williamsburg, Lee, Sumter, and Clarendon.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
6 ***** TRIPLE: council serve scout ORIGINAL: Allegheny Highlands Council serves Scouts in New York and McKean and Potter counties in Pennsylvania.	ONE NOTKIND DIST	SOME NOTKIND DIST	ONE NOTKIND DIST
7 ***** TRIPLE: disaster approach land ORIGINAL: Focuses on the process of Kundalini-Yoga, one of the stages in Aum's practice. —- Disaster Approaches the Land of the Rising Sun: Shoko Asahara's Apocalyptic Predictions, (Shizuoka: Aum, 1995).	SOME NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
8 ***** TRIPLE: cafe serve meal ORIGINAL: Supai has one small, air-conditioned lodge (Havasupai Lodge), a convenience store, and one cafe serving fast food meals.	QUANT QUANT QUANT	QUANT QUANT QUANT	QUANT QUANT QUANT
9 ***** TRIPLE: alliance plan attack ORIGINAL: With a golden opportunity too good to miss, the Alliance planned a two pronged attack.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
10 ***** TRIPLE: version double amount ORIGINAL: The Stereo 8 version doubled the amount of programming on the tape by providing eight total tracks, usually consisting of four programs of two tracks each.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
11 ***** TRIPLE: officer receive grade ORIGINAL: For example, Bavarian officers received various grades of that Kingdom's Military Merit Order (Milit r-Verdienstorden), while enlisted men received various grades of the Military Merit Cross (Milit r-Verdienstkreuz).	SOME KIND DIST	SOME NOTKIND DIST	SOME KIND DIST
12 ***** TRIPLE: club send delegation ORIGINAL: When the Jacobin Club of Mysore sent a delegation to Tippu Sultan, 500 rockets were launched as part of the gun salute.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
13 ***** TRIPLE: humanity lose insight ORIGINAL: Humanity might also lose brilliant insights gained by these new minds during the process of conceptual rediscovery.	ALL NOTKIND COLL	ONE NOTKIND DIST	ONE NOTKIND DIST
14 ***** TRIPLE: service open office ORIGINAL: The U.S. Postal Service opened the Isaacson Post Office but renamed it to Nogales in 1883.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
15 ***** TRIPLE: founder work labor ORIGINAL: The founders of Degania worked back-breaking labor attempting to rebuild what they saw as their ancestral land and to spread the social revolution.	MOST NOTKIND DIST	MOST NOTKIND COLL	MOST NOTKIND DIST
16 ***** TRIPLE: bells buy cottage ORIGINAL: The Bells later bought a cottage near Dugort and lived in it periodically until 2001 when they donated it to be used as an artists' residence.	ALL NOTKIND COLL	SOME NOTKIND COLL	ALL NOTKIND COLL

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
17 ***** TRIPLE: attack reveal key ORIGINAL: As a result, the goal of the attack are to glean enough information to weaken the scheme against a wide variety of target ciphertexts; in the most successful attack scenario, this attack might successfully reveal the secret decryption key and thus completely break the scheme.	ONE NOTKIND DIST	ALL KIND DIST	ONE NOTKIND DIST
18 ***** TRIPLE: council describe flag ORIGINAL: The Council of Europe describes the flag as: 'Against the blue sky of the Western world, the stars represent the peoples of Europe in a circle, a symbol of unity .	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
19 ***** TRIPLE: group announce deal ORIGINAL: In June 2003, Sanctuary Records group announced a deal with Morrissey.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
20 ***** TRIPLE: minister award medal ORIGINAL: In 1993, the French Minister of Culture awarded him the medal of Chevalier des Arts et des Lettres (the Order of Arts and Letters).	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
21 ***** TRIPLE: film feature song ORIGINAL: The film featured two songs by Radiohead, Fake Plastic Trees and My Iron Lung taken from the album The Bends.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
22 ***** TRIPLE: tiger select player ORIGINAL: At the draft table, the Tigers selected the following players.	MOST NOTKIND COLL	SOME NOTKIND COLL	ALL NOTKIND COLL
23 ***** TRIPLE: council serve scout ORIGINAL: Gateway Area Council serves Scouts in Wisconsin and Minnesota.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
24 ***** TRIPLE: government create reservation ORIGINAL: The government created Indian reservations for the Cahuilla in 1877.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
25 ***** TRIPLE: agreement require year ORIGINAL: This agreement required a full year of preparation on the part of the Venetians to build numerous ships and train the sailors who would man them, all the while curtailing the city's commercial activities.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
26 ***** TRIPLE: recipe take pain ORIGINAL: His recipes often take pains to demystify cooking by explaining the chemical processes at work.	MOST NOTKIND COLL	MOST KIND DIST	ALL NOTKIND DIST
27 ***** TRIPLE: effect cause consequence ORIGINAL: The ELIZA effect can also cause negative consequences if the user's assumptions do not match program behavior.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
28 ***** TRIPLE: friend get job ORIGINAL: A friend got him a job as an apprentice at a furrier.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
29 ***** TRIPLE: group join wwpdb ORIGINAL: The BMRB (USA) group joined the wwPDB in 2006.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
30 ***** TRIPLE: hormone cause apolysis ORIGINAL: This hormone causes apolysis - the separation of the cuticle from the epidermis excretion of new cuticle beneath the old degradation of the old cuticle.	ALL KIND COLL	MOST KIND COLL	ONE NOTKIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
31 ***** TRIPLE: declarer take finesse ORIGINAL: As with many deceptive plays, declarer should take the entry finesse as early in the play as possible, before the defense realizes it must play second hand high to block the suit.	ONE NOTKIND DIST	ALL KIND DIST	ALL KIND DIST
32 ***** TRIPLE: nation form alliance ORIGINAL: The World Evangelical Alliance is a network of churches in 127 nations that have each formed an evangelical alliance and over 100 international organizations joining together to give a worldwide identity, voice and platform to more than 420 million evangelical Christians'.	QUANT QUANT QUANT	QUANT QUANT QUANT	QUANT QUANT QUANT
33 ***** TRIPLE: composition include hit ORIGINAL: Blake's compositions included such hits as, 'Bandana Days', 'Charleston Rag', 'Love Will Find A Way', 'Memories of You', and 'I are Just Wild About Harry'.	ALL NOTKIND COLL	ALL NOTKIND COLL	ALL NOTKIND DIST
34 ***** TRIPLE: band release album ORIGINAL: Most recently, the band released a new studio album, Now, Diabolical.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
35 ***** TRIPLE: esterase fulfil role ORIGINAL: Different esterases fulfil this role.	SOME NOTKIND DIST	SOME NOTKIND DIST	SOME NOTKIND DIST
36 ***** TRIPLE: movie feature sequence ORIGINAL: While the martial arts movies of the 1970s generally featured highly-stylized fighting sequences in period or fantasy settings, Hanged's choreography, set in modern urban areas, was more realistic and frenetic - featuring long one-on-one fight scenes.	MOST KIND DIST	MOST NOTKIND DIST	MOST KIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
37 ***** TRIPLE: owner commission firm ORIGINAL: By 1914, the hotel's owner, Daniel White, taking a hint from the Marlborough-Blenheim, commissioned the firm of Price and McLanahan to build an even bigger hotel.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
38 ***** TRIPLE: precipitation average inch ORIGINAL: Precipitation averages 16 inches a year in this area, with snowfall of 50 inches.	ALL NOTKIND COLL	ONE NOTKIND DIST	ALL NOTKIND COLL
39 ***** TRIPLE: row show effect ORIGINAL: The first row shows the effect of the eight rotations, and the second row shows the effect of the eight reflections.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
40 ***** TRIPLE: guard say thing ORIGINAL: In a deleted scene, the Miramax security guard says the same thing before being called a dick, just as William was.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
41 ***** TRIPLE: family give land ORIGINAL: Over the years, the J. R. Phillips family has given additional land by Oscar, Horace Phillips and L. G. Phillips, and more recently Wayne Phillips.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
42 ***** TRIPLE: tune become hit ORIGINAL: Its theme tune became a huge hit.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST

Continued on next page

Table B.1 – continued from previous page

Annotation instance	A1	A2	A3
43 ***** TRIPLE: abscess require treatment ORIGINAL: Like other abscesses, perianal abscesses may require prompt medical treatment, such as an incision and debridement or lancing.	ALL KIND DIST	ALL KIND DIST	ALL KIND DIST
44 ***** TRIPLE: fiddler use replica ORIGINAL: Scottish fiddlers emulating 18th century playing styles sometimes use a replica of the type of bow used in that period, which is a few inches shorter, and weighted significantly differently.	MOST KIND DIST	SOME KIND DIST	MOST KIND DIST
45 ***** TRIPLE: resident oppose idea ORIGINAL: Some local residents also oppose the idea of funding a system they believe to be only for the benefit of out-of-town tourists.	QUANT QUANT QUANT	QUANT QUANT QUANT	QUANT QUANT QUANT
46 ***** TRIPLE: party lead front ORIGINAL: Syrian Arab Republic (Baath Party leads the National Progressive Front).	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST
47 ***** TRIPLE: facility include facility ORIGINAL: Facilities include the Travis Shooting Sports Facility, which contains an olympic shotgun shooting range, the Gates Aquatic Center, a horse corral, BB Gun range, two lakes, and the four above mentioned camps.	ALL NOTKIND COLL	ALL NOTKIND COLL	ALL NOTKIND COLL
48 ***** TRIPLE: change affect name ORIGINAL: More minor changes have affected the names of some countries named after ethnicities, whose endings have changed from -ujo to -io, and women's names ending in -a (e.g. Maria), whereas purists once insisted on using the noun ending -o (e.g. Mario or Mariino).	SOME NOTKIND DIST	SOME NOTKIND DIST	QUANT QUANT QUANT

Continued on next page

Table B.1 – continued from previous page

Annotation instance	**A1**	**A2**	**A3**
49 ***** TRIPLE: warrior finish time ORIGINAL: Once an Eldar warrior finishes his time as an Aspect Warrior, they move on to other occupations, as per the convention of the Eldar Path.	ONE NOTKIND DIST	ALL KIND DIST	ALL KIND DIST
50 ***** TRIPLE: city win lawsuit ORIGINAL: The city won the lawsuit in 1970, and the land was transferred as open space to the Golden Gate National Recreation Area.	ONE NOTKIND DIST	ONE NOTKIND DIST	ONE NOTKIND DIST