Linguistic **I**ssues in **L**anguage **T**echnology – LiLT
Volume 13, Issue 2 May 2016

Many speakers, many worlds

Interannotator variations in the quantification of feature norms

Aurélie Herbelot

Eva Maria Vecchi

Published by CSLI Publications

LiLT volume 13, issue 2 May 2016

Many speakers, many worlds

Interannotator variations in the quantification of feature norms

Aurélie Herbelot, *University of Trento*
Eva Maria Vecchi, *University of Cambridge*

1 Introduction

Quantification (see e.g. Peters and Westerståhl, 2006) is probably one of the most extensively studied phenomena in formal semantics. But because of the specific representation of meaning assumed by model-theoretic semantics (one where a true model of the world is *a priori* available), research in the area has primarily focused on one question: given a model, what does it *mean* for a speaker to utter a statement of the form $Qx[P(x)]$, where Q is a natural language quantifier such as *no, few, some, many, most, all, at least 3...* (or even a null quantifier $\emptyset$)? What is the relation of a quantifier to the truth value of a sentence? In contrast, relatively little has been said about the way the underlying model comes about, and its relation to individual speakers' conceptual knowledge.

Consider for instance the simple model in Fig.1. Given this state-of-affairs, where the set of cats and the set of black things overlap, the sentences *some cats are black* and *at least one cat is black* can be said to be logically true, while *all cats are black* is logically false. Expanding on this purely logical interpretation, researchers (e.g. Huang and Snedeker (2009), Grodner et al. (2010) and Degen and Tanenhaus (2015)) have

LiLT Volume 13, Issue 2, May 2016.
Many speakers, many worlds.

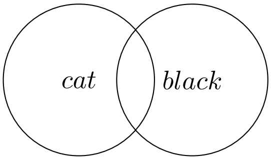

FIGURE 1 The set of cats and the set of black things overlap.

further shown how quantifier preferences depend on a variety of factors ranging from the cardinality of the set involved to the 'Question Under Discussion' in the discourse. That is, depending on the number of cats involved in the state-of-affairs under consideration, humans may prefer the description *two cats are black* to *some cats are black*, despite both sentences being true.

What semantics and pragmatics have so far failed to provide is an account of models themselves: how is it that in the first place, a speaker might model the world in a way that the set of cats and the set of black things overlap? In grounded situations, perceptual input arguably constrains the beliefs of the speaker about the observed state-of-affairs. If we stand in a room with five cats, two of which are black, it seems intuitive to infer that our model of the state-of-affairs includes a set of five cats, and the overlap between the set cats and the set of black things has cardinality 2. It is much less clear how the model comes about in non-grounded situations (which constitute the majority of an adult's utterances). Having encountered only a small proportion of all cats in the world, and perhaps no unicorn, why is it that we confidently utter sentences such as *most cats have four legs* or *all unicorns look like horses*, pointing at an underlying model where indeed, most cats have four legs and all unicorns look like horses?

In this paper, we make a first step in investigating how native speakers of English model relations between non-grounded sets, by observing how they quantify simple statements. For instance, we ask how an individual might quantify *bats are blind* (some? all?), hoping to gain a representation of their underlying model of bats (and blindness). Note that explicit quantification is an unusual phenomenon in that it cannot directly be studied from corpora, being relatively rare in naturally occurring text: underspecified constructions like bare plurals and (in)definites starting with *a/the* are much more frequent than the equivalent *some/most/all*-quantified NPs.[1] Herbelot and Copestake (2011) estimate that around 7% of noun phrases are explicitly quantified. This means that we are unlikely to find out from a corpus study, for instance, that *all* cats are mammals: the generic *cats are mammals* is the standard way to express the predication. Our main contribution is to remedy this lack of data by releasing an annotation layer for a well-known set of feature norms (the 'McRae norms', McRae et al., 2005) consisting of over 7,000 concept-feature pairs, labelled by 3 native speakers of English. For each pair in the norms, coders have provided a natural language quantifier, resulting in new statements such as *all tricycles have three wheels* or *few apes are blind* to extend the data for pertinent computational tasks.

This paper is structured as follows. We first give some motivation for our task, from both a theoretical linguistic and computational semantic point of view (§2). We then describe our annotation setup (§3) and follow on with an analysis of the produced dataset, conducting a quantitative evaluation which includes inter-annotator agreement for different classes of predicates (§4). We observe that there is significant agreement between speakers but also noticeable variations. We posit that in set-theoretic terms, there are as many worlds as there are speakers (§5), but the overwhelming use of underspecified quantification in ordinary language covers up the individual differences that might otherwise be observed.

2 Motivation

Although quantification is rarely explicit in naturally occurring text, it is intrinsic to most utterances. Any reference act picks out some set of individuals X in a world and, by associating a predicate P with it, builds a model which is interpretable in terms of a quantified relation: some, most, all individuals in X do P. This process happens intuitively so that, when someone utters *Mice are in the cellar*, we don't assume that all mice in the world have gathered in the speaker's cellar, only *some* of them. In this paper, we will regard this process as generating a natural language quantifier with a set-theoretic interpretation. Note that we are not making any claims about whether speakers have set-theoretic models 'in their head'. We only wish to argue that their interpretation of a statement involves building *some kind of* model of the state-of-affairs described by the speaker, and that this model gives them some (rough) information about the proportion of instances of a

[1] There may be certain genres where there are more or less uses of quantifiers, however it still remains rare in ordinary natural language text.

concept involved in the situation. We use set theory as a shorthand to express the end-product of this intuitive process.

Being able to generate a quantifier for a given subject-predicate pair is a prerequisite for many lexical semantics and inference tasks. It is arguably part of the concept *cat* that some cat instances are tabbies and all are mammals, and knowing this allows us to infer that given a cat taken at random, that cat is most definitely a mammal and may be a tabby. A lot of work in computational semantics has focused on extracting specific set relations from text, in particular those involving set identity or set inclusion (e.g. synonymy, hyponymy: Landauer and Dumais, 1997, Hearst, 1992 through to Bullinaria and Levy, 2012, Baroni et al., 2012, Lenci and Benotto, 2012). But much less research has looked into the problem of generally inducing a set-theoretic model from corpus data, including set relations which can range from small to consequent – but not necessarily universal – overlap (Herbelot, 2013, Herbelot and Vecchi, 2015).

The dataset we release with this paper has two motivations. The first is to gather linguistic data to help us understand, from a theoretical point of view, to which extent humans agree on a single model of the world. The second is to provide a large gold standard of quantified predications which can be used as training/test data in computational tasks such as entailment, inference, concept modelling, etc.

3 Quantifying the McRae norms

The McRae norms (2005) are a set of feature norms elicited from 725 human participants for 541 concepts. The annotators were asked to provide features for each concept, covering physical, functional and other properties. The result is a set of 7257 concept-feature pairs such as *airplane used-for-passengers* or *bear is-brown*.

We conducted the quantification of the McRae data in the following way. We recruited three native English speakers (one Southeast-Asian and two American speakers, henceforth denoted as *A1* and *A2/A3*), all computer science students. For each concept-feature pair (C, f) in the norms, they were asked to provide a label expressing the ratio of instances of C having the feature f. The allowable classes were NO, FEW, SOME, MOST, ALL. Table 1 provides example annotations for concept-feature pairs. An additional label, KIND, was introduced for usages of the concept as a kind, where quantification does not apply (e.g. *beaver symbol-of-Canada*).

Our investigation required minimising the pragmatic interferences observed in quantifier selection. Note that the way a speaker quantifies

Concept	*Feature*	
ape	is_muscular	ALL
	is_wooly	MOST
	lives_on_coasts	SOME
	is_blind	FEW
tricycle	has_3_wheels	ALL
	used_by_children	MOST
	is_small	SOME
	used_for_transportation	FEW
	a_bike	NO

TABLE 1 Example annotations for concepts.

bats are blind depends on a) the speaker's beliefs about the concepts *bat* and *blind* and b) their personal interpretation of quantifiers in context. The first aspect concerns matters of lexical semantics and, broadly-speaking, world knowledge. Does the speaker understand blindness as complete lack of sight or (just) poor sight? What do they know about bats? The second aspect relates to the pragmatics of quantifier semantics: we straightforwardly observe, for example, that *all* has a much wider meaning than ∀ suggests (as in *all my friends say I'm right*, which typically does not imply universal quantification). These two aspects had to be clearly separated in our study, as we focused on the first question, i.e. what people believe about the actual state of the world (regardless of their way of expressing it), and how this relates to their conceptual and lexical knowledge.

In order to reduce such interferences, we gave clear instructions to the coders on how to use the annotation labels (reproduced in the Appendix). We defined the label ALL as a 'true universal' which either a) doesn't allow exceptions (as in the pair *cat is-mammal*) or b) may allow some conceivable but 'unheard-of' exceptions. In other words, we wanted ALL to refer to near-definitional features and tried to prevent participants from worrying about far-fetched exceptions to the norm. The label MOST was used for all majority cases, including those where the annotator knew of actual real-world exceptions to a near-definitional norm. The NO/FEW distinction was defined as mirroring ALL/MOST. SOME was not associated with any specific instructions.

To further minimise potential disagreements, we introduced extra instructions for cases where the participants might hesitate between two labels. We encouraged them to choose the label corresponding to lower set overlap (i.e. prefer SOME to MOST, MOST to ALL, etc). This ordering preference was set up with a view to use the dataset in computational inference tasks, where truth preservation is important: while *cats have four legs* should ideally be annotated as MOST, the label SOME still results in a true predication (*some cats have four legs*). Compare with *cats are black*, where choosing MOST over SOME would result in a false sentence (*most cats are black*).

Clearly, fixing the interpretation of the labels affects the type of information encoded by the dataset. We are *not* modelling the way that speakers naturally use the determiners *some, most, all*, etc. Rather, we are modelling the perceived overlap between the set denoted by a noun and the set denoted by a predicate (simplifying somewhat and regarding predicates as sets rather than functions). Our use of a script font for our allowable labels (e.g. SOME, MOST) reflects the fact that we have fixed the meaning of the corresponding quantifier in one of their possible interpretations. Fixing the labels' interpretation, of course, does not completely suppress all unwanted effects. For instance, understanding ALL as a clear universal does not prevent annotators from falling into the 'generic' trap identified by Leslie et al. (2011), whereby people often agree to false statements such as *all ducks lay eggs* due to the straightforward availability of the corresponding generics (*ducks lay eggs*).

The order of the data was randomised prior to distributing it to the annotators. Participants took 20 or less hours to complete the task, which they did at their own pace, in as many sessions as they wished. While the task was a significant time investment, having a full set of labels for each participant allowed us to compare agreement over different classes of predicates (§4.3) and thus obtain an insight into significant variations at the individual level.

4 Data analysis

This section describes the annotated dataset, concentrating on three aspects: the overall distribution of the six labels, the overall inter-annotator agreement, and specific variations in agreement across conceptual feature classes.

4.1 Class distribution

Fig. 2 shows how the general distribution of the annotation varies across participants. As we might expect, the labels KIND and NO are seldom used: this can be easily explained by noting that KIND mentions are overall rare, and that the feature norms should by definition apply to the concept under consideration.

As far as the other quantifiers are concerned, we note relatively wide variations across annotators. *A1*, in particular, uses ALL extensively,

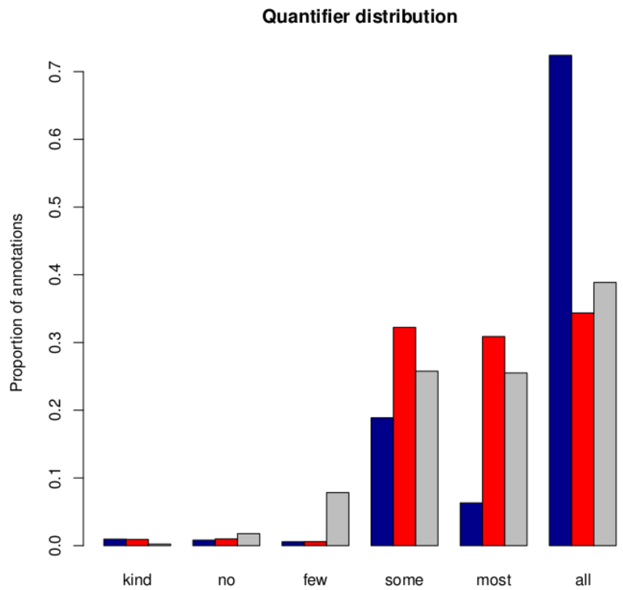

FIGURE 2 Class distribution per annotator ($A1$: blue, $A2$: red, $A3$: grey).

applying the label to over 70% of the McRae instances. A manual analysis of the data reveals that the annotator may have interpreted their model of the world in a much more normative way than the other participants. For instance, they may have labelled the pair *ambulance – used-for-rescuing* with a universal under the assumption that the intrinsic function of an ambulance is to rescue people, regardless of the fact that some ambulances might finish their days as museum objects or converted vehicles. The generalisation effect noted by Leslie et al. (2011) (e.g. universalising *ducks lay eggs*) may also be at fault, but it is impossible to tell to what extent this might be the case. The distributions of *A2* and *A3* are much more alike – although smaller variations can be found between them too. Notably, *A3* uses FEW significantly more than the other two participants.

As we will show in 4.3, looking at the data in more detail also reveals that the overall similarity between *A2* and *A3* does not hold across all categories of predicates. In fact, some categories show lower agreement between *A2* and *A3* than between either *A2* or *A3* and *A1*.

Having analysed the overall distribution of all annotations, we remove the instances marked with at least one KIND label, which potentially lack a quantificational interpretation. We end up with a remaining 7138 instances out of 7257.

4.2 Inter-annotator agreement

Given the differences observed in the use of each individual quantifier, we need an inter-annotator agreement measure that assumes separate distributions for all three coders. We would also like to account for the seriousness of the disagreements: for instance, a disagreement between NO and ALL should be penalised more than one between MOST and ALL. We select weighted Kappa (κ_w) (Cohen, 1968) as our agreement measure, since it satisfies both requirements. κ_w is a variant on the kappa inter-annotator agreement measure. Simple kappa (Cohen, 1960) measures the extent to which two annotators agree above what would be expected by chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{1.1}$$

where p_o is the observed agreement between annotators and p_e the expected agreement.

In this version, all disagreements are weighted the same. The weighted version of kappa, in contrast, allows for penalising different types of disagreement in different measures. The weights are given by an $k * k$ matrix, where k is the number of classes under consideration. The weighted kappa for two annotators is given by:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} o_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} e_{ij}} \tag{1.2}$$

where w_{ij} is the weight assigned to a disagreement involving classes i and j, and o_{ij} and e_{ij} are respectively the observed and expected agreements for that pair of classes.

As κ_w can only be calculated for two annotators, we report all annotator pairs κ_w^{12}, κ_w^{13} and κ_w^{23}, as well as their average (κ_w^{A}), computed using the R 'psych' package.[2]

Calculating κ_w requires setting a weight matrix to control the penalty applied to specific disagreements. Ideally, we would like this weight matrix to reflect the prevalence of the predication (i.e. the set-theoretic ratio between the restrictor and scope of the quantifier). So in a world where MOST corresponds to around 80% of instances of C having property f and ALL 100%, the penalty for a confusion between MOST and ALL should be set to $100 - 80 = 20$.

[2]http://cran.r-project.org/web/packages/psych/psych.pdf

Predication type	Example	Prevalence
Principled	Dogs have tails	92%
Quasi-definitional	Triangles have three sides	92%
Majority	Cars have radios	70%
Minority characteristic	Lions have manes	64%
High-prevalence	Canadians are right-handed	60%
Striking	Pit bulls maul children	33%
Low-prevalence	Rooms are round	17%
False-as-existentials	Sharks have wings	5%

TABLE 2 Classes of generic statements with associated prevalence, as per Khemlani et al. (2009).

Quantifiers are however notoriously difficult to associate with stable prevalence estimates (i.e. ALL might correspond to 90%, 95%, 100% of a set, depending on its context of use). Even in our scenario, where we are forcing a certain interpretation of the quantifiers, MOST could easily range from, say, 80% to 99%. The best we can do is to provide a mean for each quantifier, so that, for instance, $Pr(\text{SOME})$ is the average prevalence of all predications annotated with SOME. We consider two ways to get such averages:

1. Prevalence estimates have been previously elicited in Khemlani et al. (2009) (henceforth $KH09$), where 50 generic predications received an estimate from 17 coders. We use the results of this study to set κ_w's weight matrix.
   $KH09$ did not work on quantifiers *per se* but on types of generic statements, so their proposed classification must be mapped to ours for comparison. We reproduce their results in Table 2, including an example of each class, as included in their original paper. The 'quasi-definitional' class clearly corresponds to an ALL quantification, while the 'false-as-existential' corresponds to NO. So we give a prevalence of 92% to ALL and of 5% to NO. Similarly, the low-prevalence class can be mapped onto FEW, as it refers to predicates which are existentially true for a small number of instances. We also average the 'striking' and 'minority characteristic' class to get a prevalence for SOME (49% – probably an overestimate, as the minority characteric class tends to elicit inflated prevalences). We finally conflate the 'high prevalence', 'majority' and 'quasi-definitional' generics to obtaining an average prevalence of 74% for MOST.

2. We also exhaustively try all possible prevalence values in the

	Best	*KH09*
NO	0	5
FEW	5	17
SOME	35	49
MOST	95	74
ALL	100	92

TABLE 3 Prevalence estimates for each class. *Best* shows the estimates that led to the highest κ_w, reported with those derived from $KH09$.

	κ_w^{12}	κ_w^{13}	κ_w^{23}	κ_w^{A}
full				
KH09	.37	.34	.50	.40
BEST	.44	.40	.50	**.45**
maj				
KH09	.49	.48	.60	.52
BEST	.57	.53	.67	**.59**

TABLE 4 κ_w for MCRAE$_{full}$ and MCRAE$_{maj}$.

range 0-100, with the only constraint that Pr(NO) < Pr(FEW) < Pr(SOME) < Pr(MOST) < Pr(ALL). We record κ_w^A for each combination, hoping to find that the best agreement does roughly correspond to the prevalence values elicited by $KH09$.

We calculate κ_w on our full dataset (denoted here as MCRAE$_{full}$), as well as on the subset in which there was majority agreement among annotators (i.e. where two or more annotators used the same label: MCRAE$_{maj}$, 6120 instances, corresponding to 86% of our data). MCRAE$_{maj}$ can straightforwardly be turned into a gold standard for any computational task by setting the quantification of each instance to the majority class. Table 3 reproduces the prevalences derived from $KH09$, alongside the estimates that led to the highest κ_w overall in the systematic search (marked *Best*). Table 4 reports the calculated kappa values for both MCRAE$_{full}$ and MCRAE$_{maj}$.

We find that κ_w^{23} is consistently higher than κ_w^{12} and κ_w^{13}, indicating better agreement between $A2$ and $A3$. This is expected given the differences in class distributions observed in Fig. 2. The $KH09$ estimates give reasonable kappas, reaching 0.52 for MCRAE$_{maj}$. But a significant improvement in agreement can be observed when systematically searching for the 'best' weight matrix (κ_w^A=0.59 for MCRAE$_{maj}$). The corresponding prevalences show MOST and ALL, as well as NO and FEW, to be virtually indistinguishable.

These results indicate that, as far as prevalence was concerned, our annotators interpreted MOST as a near-universal, probably analogous to the 'principled' class in $KH09$ – even though our guidelines would have left some scope for a less strict majority reading. For some applications, users of the dataset may thus want to conflate the MOST and ALL classes. However, we also note that out of the 6120 instances in MCRAE$_{maj}$, 1136 correspond to a majority of MOST annotations –

giving some sizeable data for the comparison of universals and near-universals.

Finally, we consider the correlation between the original production frequencies and the annotation agreement for each concept-feature pair. The production frequency of a feature for a concept is the number of times it was generated in the original McRae experiment. For example, the property *is-crunchy* was produced 11 times for the concept *apple*. In doing this, we test whether a feature that is very salient for a concept leads to a more stable set relation across speakers. We first compare the amount of agreement among annotators (0:no agreement; 1:majority agreement without consensus; 2:unanimous consensus) and the original production frequencies: this results in a very low correlation (Spearman's ρ <0.2). This tells us that high agreement values can be expected in cases of high production frequency, as well as cases of very low production frequency. Indeed, *is-yellow* may be produced with high frequency for *banana* and still not prevent annotators from interpreting the concept as referring to either *all* bananas or only ripe ones. Conversely, few people may produce *an-inanimate* for *anchor*, but the relevant set relation is unarguably one of inclusion.

We then attempt to test the correlation between the prevalence estimates of quantifiers and the original McRae production frequencies to see if there is a direct relationship between the production of a feature and the proportion of instances having that feature (using the majority opinion from MCRAE_{maj}). The assumption here is that a feature that is shared by all instances of a concept is more likely to be produced. Again, we obtain very low correlation (Spearman's ρ <0.3). This result underlines the fact that we cannot extract or estimate quantifier values directly from the feature norms. Instead, it is clear that we need a dataset where that information is explicitly annotated.

4.3 Analysis of various feature types

The McRae norms are annotated with feature classes which correspond to types of knowledge stored in separate brain regions (marked as 'BR Features' in the data – see Cree and McRae, 2003). These classes map onto different modalities such as colour, shape or taste, which have been found to activate particular areas of the brain. This includes 'function' for predicates denoting the use of an object (e.g. *hoe used-for-farming*), or 'tactile' for features associated with the sense of touch (e.g. *toaster is-hot*). In addition, two categories were defined for features that fell out of the brain region classification: the 'taxonomic' category for *is-a* relations (e.g. *axe is-a-tool*), and the 'encyclopedic' category, designed as a catchall for all other features. Table 5 shows the different classes, together with examples of corresponding predications. It also records the frequency of each class in our data (after the instances marked KIND were removed), and the inter-annotator agreements (pairs and average, using the *Best* weights obtained in 4.2).[3]

The agreement results show several interesting effects. First, while we noticed that overall, $A2$ and $A3$ agreed significantly more than $A1$ with either of them, it turns out that for specific feature classes, this tendency does not hold. For instance, $A1$ and $A2$ obtain much better agreement on 'visual-colour' items than either with $A3$. This is also the case for the 'taxonomic' class. This result indicates that, as we might expect, differences in human perception and conceptual make-up are reflected in their use of quantifiers. Note that this does not seem to be linked to cultural effects: the Southeast-Asian speaker ($A1$) and one of the American speakers ($A2$) seem to have a closer notion of colour than the two Americans ($A2/A3$). Similarly, $A1$ and $A3$ share much better agreement on 'smell' features than $A2$ and $A3$ – pointing at individual rather than cultural differences.

Second, the ranking of classes by κ_w^A highlights several notable facts. One is that, although at the top of the table, the 'taxonomy' class does not result in as good an agreement as we might expect. Annotators disagreed on examples such as *bulls are cows*, *cats are pets*, or again *cloaks are coats*. While the second of those examples probably does relate to actual disagreements in quantification, the other two seem to be artefacts of conceptual differences: what are cows, cloaks and coats? Or in other words, which individuals should be included in the sets of cows, cloaks and coats?

Another enlightening aspect is the kappa values obtained by different types of perceptual classes. While the 'form and surface' class comes in second position in the ranking, 'colour' and 'motion' features get much lower kappas. Perhaps expectedly, 'smell', 'taste', 'tactile' and 'sound' features are at the bottom of the table: these features correspond to senses that are on the whole less emphasised in English.

Generally, the observed ranking may be explained by the type of cognitive process at work in the quantification task. We note that there is evidence for quantification being relatively straightforward in

[3]The R psych package does not calculate kappa in cases where the contingency table is unbalanced – i.e. whenever annotators did not use the same set of labels. Because of this, we encountered problems when calculating κ_w for the three classes 'smell', 'taste' and 'tactile' (marked by an asterisk in Table 5), as the NO and FEW quantifiers had only been used by one annotator. In order to overcome this issue, we made two minor changes to each of these files, changing one ratings from FEW to NO, and one from SOME to FEW.

BR Label	Example	Freq.	κ_w^{12}	κ_w^{13}	κ_w^{23}	κ_w^{A}
taxonomic	axe a_tool	713	.66	.48	.56	.57
visual-form	ball is_round	2330	.48	.44	.54	.49
function	hoe used_for_farming	1489	.36	.35	.50	.40
encyclopaedic	wasp builds_nests	1361	.39	.34	.37	.37
visual-colour	pen is_red	421	.44	.27	.30	.34
visual-motion	canoe floats	332	.28	.20	.46	.31
∗smell	skunk smells_bad	24	.34	.48	.12	.31
∗taste	pear tastes_sweet	84	.22	.29	.36	.29
∗tactile	toaster is _hot	242	.19	.31	.30	.27
sound	tuba is_loud	143	.11	.10	.36	.19

TABLE 5 Per-feature agreement for MCRAE$_{full}$, sorted by κ_w^{A}

some grounded contexts (those involving exact, rather than approximate number sense, and small cardinality – see Clark and Grossman, 2007). But quantifying feature norms involves using one's approximate number sense over large, non-grounded sets. This is bound to affect agreement for non-definitional features, i.e. those contingent features which cannot be abstractly derived (see *bottle is-green* vs. *axe is-tool*).

When looking more closely at the data, it seems clear that vague and gradable adjectives affect agreement negatively. This explains the relatively low kappa for the 'colour' class, as well as the four lowest classes in the table. For example, the 'sound' class contains a significant proportion of features such as *is-loud, is-quiet, produces-high-pitched-sounds*, etc. However, this is not the only issue. It seems that in many cases, a statement was read by an annotator as involving some kind of potentiality, and labelled accordingly. For instance, *missile explodes* received the labels SOME, MOST and ALL. It is likely that the SOME interpretation quantifies over missiles which actually explode, while the MOST/ALL interpretation considers the potential of a missile to explode. A similar explanation can be provided for predications such as *mouse squeaks* or *balloon floats*.

Overall, this short analysis illustrates that, even when features are reliably produced for a given concept, their quantification may vary significantly between annotators and agreement is highly dependent on the corresponding functional or sensory type. To finish this section, we will come back to our initial observation that most NPs in English lack explicit quantification (see §1). In addition, we should mention that several studies have shown that generics are acquired by children much earlier than quantifiers (e.g. Hollander et al., 2002), strengthening arguments for the general precedence of 'vague' quantification. Given the results reported here, it seems fair to assume that communication is generally more successful when avoiding explicit quantification: a speaker is more efficient in uttering *tubas are loud* than the potentially controversial *some tubas are loud* (unless, for pragmatic reasons, they want to emphasize the quantification). This remark also holds for readings of adverbial or modal quantification which involve quantification over individuals, as in *tubas are sometimes loud*, *tubas can be loud*.[4] This is not to say that speakers' models of the world are fundamentally different: the weighted kappas obtained on our dataset are very reasonable, and in over 86% of the 7,138 instances 2 or 3 annotators agreed on their judgment (majority agreement). Still, disagreements are common enough that they might be costly in conversation.

5 Conclusion

In this paper, we have presented an annotation layer for the McRae feature norms (McRae et al., 2005), which shows quantifier labels of each concept-feature pair in the norms, as given by three native speakers of English. We are freely releasing this data for future research.[5] A subset of the dataset totalling 6120 instances contains all cases of majority agreement and can easily be used as gold standard for any computational application requiring examples of explicitly quantified statements about a range of concepts.

For evaluation purposes, we systematically matched the quantifiers under study to a range of prevalence estimates and calculated the corresponding weighted kappas over our data. Following the assumption that more accurate estimates should result in better kappa agreement, we derived prevalence figures for each one of our five quantifiers. For MOST, we found that, within the scope allowed by the guidelines (simple majority to near-universal), annotators applied the label to cases with prevalence close (but not equal) to 100%, clearly reading the quantifier as a 'universal with exceptions'. In spite of this, SOME did not end up covering cases of 'logical' majority (i.e. anything over 50%): with a prevalence of 35%, it was interpreted as a simple existential. This result is interesting, as it suggests that coders may not have felt the need for a quantificational category covering generally high prevalence. We also showed that agreement is not correlated with the frequency

[4] For some discussion on the relation between event and individual quantification in generic sentences, see Dobrovie-Sorin (2003).

[5] The full dataset, the majority cases and our annotation script are available at `http://www.aurelieherbelot.net/research/computational-linguistics-resources/`.

of feature production, indicating that a feature which is widely seen as relevant for a concept may still cause disagreements with regard to the set-theoretic interpretation of the norm. Finally, we observed that inter-annotator agreement was strongly dependent on the type of feature involved, with non-visual, sensory features generating more disagreements than definitional or functional features.

While overall, we observe good agreement on our quantification task (reaching κ_w^A=.59 for instances with a majority opinion), it seems unwarranted to assume that generalised quantifiers are entirely and reliably shared amongst speakers. Rather, we must posit a 'many speakers, many worlds' hypothesis: individuals share some generic conceptual knowledge which helps them efficiently communicate, but their actual models of the world (what exactly counts as a cloak, and which proportion of those are coats) can differ fairly significantly. In addition, it is unclear how cognitively complex the quantification process is for a given speaker. While we did not monitor the difficulty of our task per se, it seems fair to assume that it is non-trivial: a simple pair such as *bowl used-for-eating* requires retrieving all subkinds of bowls the speaker may be familiar with, building a model of the entire set (across subkinds) and finally making a decision about the quantifier. It is not unreasonable to think that this process is only called upon when absolutely needed.

This has consequences for the way we formalise models and, more specifically, quantification and inference. Our observations above point at models which, most of the time, are left underspecified but *can* be specified when necessary – although at some cost and with some inter-speaker variations (for an example of a formalisation of underspecified quantification, see Herbelot and Copestake, 2011). Such variations will presumably affect agreement in logical inference tasks. For instance, inferring the probable colour of a hypothetical bathtub may turn out to be non-trivial: the fact that speakers produce the norm *is-white* for the corresponding concept may not be correlated with any expectation with regard to individuals (leading to the three annotations SOME, MOST and ALL in our data).

We conclude by arguing in favour of a speaker-dependent notion of model which accommodates variations in people's beliefs about the world, while satisfying the requirement for broad general agreement, necessary for successful communication. We hope, at any rate, that the dataset we are releasing will be of use for further investigations of this question.

Acknowledgments

Eva Maria Vecchi is supported by ERC Starting Grant DisCoTex (306920). The dataset creation was made possible by the Alexander von Humboldt Foundation, as part of a Postdoctoral Research Fellowship to Aurélie Herbelot. We thank our annotators for their work, as well as the anonymous reviewers for their thorough comments.

References

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the fifteenth Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pages 23–32. http://anthology.aclweb.org/E/E12/E12-1004.pdf.

Bullinaria, John A and Joseph P Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods* 44(3):890–907. http://dx.doi.org/10.3758/BF03193020.

Clark, Robin and Murray Grossman. 2007. Number sense and quantifier interpretation. *Topoi* 26(1):51–62. http://dx.doi.org/10.1007/s11245-006-9008-2.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46. http://psycnet.apa.org/doi/10.1177/001316446002000104.

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213. http://psycnet.apa.org/doi/10.1037/h0026256.

Cree, George S and Ken McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General* 132(2):163. http://psycnet.apa.org/doi/10.1037/0096-3445.132.2.163.

Degen, Judith and Michael K Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science* 39(4):667–710. http://onlinelibrary.wiley.com/doi/10.1111/cogs.12171/full.

Dobrovie-Sorin, Carmen. 2003. Adverbs of quantification and genericity. *Empirical Issues in Formal Syntax and Semantics* 4:27–42.

Grodner, Daniel J, Natalie M Klein, Kathleen M Carbary, and Michael K Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1):42–55. http://dx.doi.org/10.1016/j.cognition.2010.03.014.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING92)*, pages 539–545. Nantes, France. http://www.anthology.aclweb.org/C/C92/C92-2082.pdf.

Herbelot, Aurélie. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*. Potsdam, Germany. https://www.aclweb.org/anthology/W/W13/W13-0204.pdf.

Herbelot, Aurélie and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*. Oxford, England, UK. https://aclweb.org/anthology/W/W11/W11-0118.pdf.

Herbelot, Aurélie and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal. https://www.aclweb.org/anthology/W/W13/W13-0204.pdf.

Hollander, Michelle A, Susan A Gelman, and Jon Star. 2002. Children's interpretation of generic noun phrases. *Developmental Psychology* 38(6):883. http://psycnet.apa.org/doi/10.1037/0012-1649.38.6.883.

Huang, Yi Ting and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology* 58(3):376–415. http://dx.doi.org/10.1016/j.cogpsych.2008.09.001.

Khemlani, Sangeet, Sarah-Jane Leslie, and Sam Glucksberg. 2009. Generics, prevalence, and default inferences. In *Proceedings of the 31st annual conference of the Cognitive Science Society*, pages 443–448. Cognitive Science Society Austin, TX. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.7341&rep=rep1&type=pdf.

Landauer, Thomas K and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211–240. http://psycnet.apa.org/doi/10.1037/0033-295X.104.2.211.

Lenci, Alessandro and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM2012)*, pages 75–79. http://www.aclweb.org/old_anthology/S/S12/S12-1012.pdf.

Leslie, Sarah-Jane, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language* 65(1):15–31. http://dx.doi.org/10.1016/j.jml.2010.12.005.

McRae, Ken, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods* 37(4):547–559. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.6986&rep=rep1&type=pdf.

Peters, Stanley and Dag Westerståhl. 2006. *Quantifiers in language and logic*. Oxford University Press.

Appendix: Annotation guidelines for the McRae quantification task

You have been given a text file containing concept-feature pairs. The features associated with each concept are things that some people might judge salient for that concept. For instance, some people strongly associate 'made_of_wax' with 'candle'.

For each concept-feature pair, your task is to decide which proportion of the things designated by that concept actually share the feature associated with it *in the real world*. For example, you might decide that in the real world, 'all' candles are made of wax, or again that 'most' tables have four legs. We will call this 'quantifying' the concept-feature pair.

You can quantify each pair using any of the following labels:

- **all**: a universal. This applies to 'truly' universal features, i.e. those that do not accept exceptions (e.g. 'mammal' for 'cat'). It also applies to features which are *nearly* universal, i.e. features which you can conceive might be missing in some instances of the concept, but without having ever heard of such a case. So you might decide, for instance, that it is conceivable for a cat to be born without eyes, but have never heard of this happening. In that case, you would quantify the pair 'cat has_eyes' with *all*.
- **most**: majority case (e.g. 'has_4_legs' for 'cat'). This also applies to cases where exceptions are conceivable and known of (e.g. 'is_black' for 'raven': you might know that a small quantity of ravens are albinos).
- **some**: self-explanatory.
- **few**: applies to (conceivable and known of) exceptions (e.g. Few ravens are albinos).
- **no**: negated universal (e.g. the feature 'fish' for the concept 'cat').
- **kind**: this applies to cases where the feature does not relate to instances of the concept but to the concept itself. For instance, 'on_Lebanese_flag' might be a feature of 'cedar_tree', but it does not apply to individual trees, just to the concept itself.

Extra guidance

- In case of doubt, select the 'weaker' quantifier (*most* has precedence over *all*, *some* over *most*, etc.)
- There is no right answer, the most important aspect of the task is consistency, so just use your intuition to complete it. But if you really get stuck, you may look for information using an external resource (Internet, encyclopedia, etc.)