*L*inguistic *I*ssues in *L*anguage *T*echnology – *LiLT*
Volume 2, Issue 4 March 2015

From concepts to models: some issues in quantifying feature norms

Published by CSLI Publications

From concepts to models: some issues in quantifying feature norms

1.1 Introduction

Quantification (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 $\emptyset$)?

In contrast, relatively little has been said about the way individual speakers select quantifiers in particular sentential contexts. For instance, can we predict how a native speaker of English might quantify *bats are blind*? (Some? All?) The answer to this question 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 is arguably a matter of lexical semantics and, broadly-speaking, world knowledge. The second aspect relates to the pragmatics of quantifier semantics: we straightforwardly observe, for example, that *all* has a much wider meaning than $\forall$ suggests (as in *all my friends say I'm right*, which typically does not imply universal quantification – see Lasersohn, 1999 for a related discussion).

From a computational point of view, quantifier selection is an un-

LiLT Volume 2, Issue 4, March 2015.
From concepts to models: some issues in quantifying feature norms.
Copyright © 2015, CSLI Publications.

usual phenomenon in that it cannot be studied directly from corpora. The reason for this is that explicitly quantified noun phrases are very 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] So 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.

In this paper, we take a major step in the investigation of quantifier selection by producing a large-scale dataset of quantified predications. We describe 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 unattested statements such as *all tricycles have three wheels* or *few apes are blind*. This effort can be regarded as annotating conceptual knowledge (feature norms) with model-theoretic information (an indication of the set overlap between a concept and a feature). We conduct a quantitative evaluation of the dataset, including inter-annotator agreement for different classes of features, and draw some preliminary conclusions with regard to the relation between conceptual and set-theoretic apparatuses.

1.2 Motivation

Although quantification is rarely explicit in naturally occurring text, it is intrinsic to most utterances. Any statement performing reference 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 *Cats are in my garden*, we don't assume that all cats in the world have gathered in the speaker's garden, only *some* of them. The ubiquity of quantification suggests that there is a need to be able to model this information across computational tasks.

Being able to generate a quantifier for a given subject-predicate pair is in particular a prerequisite for many lexical semantics and inference tasks. 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

[1] Herbelot and Copestake (2011) estimate that around 7% of noun phrases are explicitly quantified.

Dumais, 1997, Hearst, 1992 through to Bullinaria and Levy, 2012, Baroni et al., 2012). But arguably, the whole range of possible set overlaps, from inclusion to disjointness, is necessary to fully define a concept – e.g. *all cats are mammals*, *most cats have four legs*, *some cats are black*, *no cats fly*. Similarly, explicit quantification helps deriving logically entailed sentences, including probabilistic information, for a statement. For instance, *most cats have four legs* logically entails both *some cats have four legs* and *it is likely that Sandy's new cat has four legs* (and affords many more pragmatic inferences).

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, how humans perform quantifier selection. 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.

1.2.1 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 (two American and one Southeast-Asian speakers), all computer science students. For each concept-feature pair (C, f) in the norms, they were asked to provide a natural language quantifier expressing the ratio of instances of C having the feature f. The allowable quantifiers 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*).

As pointed out in §1.1, the quantifier selection process is dependent on both the meaning attributed to the concepts involved in the predication and the meaning attributed to quantifiers themselves. The quantification of *bats are blind* may vary according to the interpretation of *blind* (complete lack of sight vs. poor sight) and world knowledge (the speaker may truly believe that bats lack a sense of sight). Similarly, *all* may be interpreted as '*every single one*' in *all cats are mammals*, '*all normal*' in *all dogs have four legs*, or again '*the great majority*' in *all my friends agree with me*. Leslie et al. (2011) have shown that people will even agree to the false statement *all ducks lay eggs* due to the straight-

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.

forward availability of the corresponding generic *ducks lay eggs*. Given such complexity, we needed to restrict the scope of interpretations for at least one aspect of the process.

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.

Participants took 20 or less hours to complete the task, which they did at their own pace.

1.3 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.

1.3.1 Class distribution

Fig. 1 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

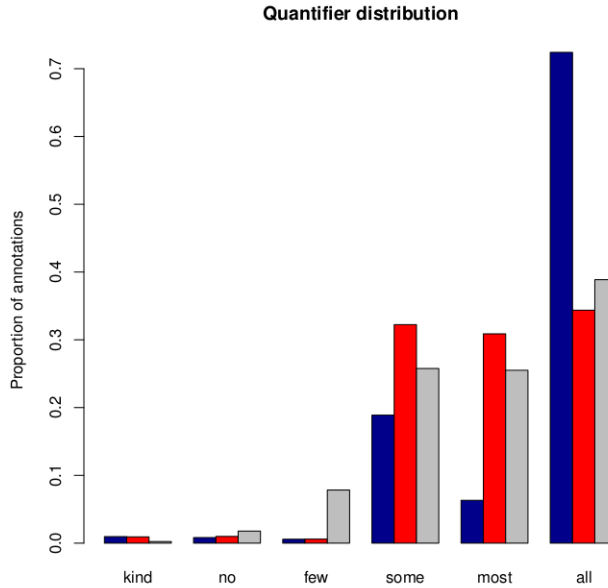

FIGURE 1 Class distribution per annotator ($A1$: blue, $A2$: red, $A3$: grey).

variations across annotators. *A1*, in particular, uses ALL extensively, applying the label to over 70% of the McRae instances. This could either be due to the generalisation effect noted by Leslie et al. (2011) or to a language variety effect: the distributions corresponding to the two American speakers (*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.

1.3.2 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. 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]

[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).

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 $100 - 80 = 20$.

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). 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. Such averages 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.

	Best	$KH09$
NO	0	5
FEW	5	17
SOME	35	49
MANY	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$_{maj}$.

As an additional check, we also exhaustively try all possible prevalence values in the range 0-100, with the only constraint that $Pr(\text{NO}) < Pr(\text{FEW}) < Pr(\text{SOME}) < Pr(\text{MOST}) < Pr(\text{ALL})$. 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 the full set of McRae norms (denoted here as MCRAE$_{full}$), as well as on the subset in which there was majority agreement among annotators (i.e. two or more annotators used the same label: MCRAE$_{maj}$, 6120 instances). 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. 1. 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 $KH09$. 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. 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.

1.3.3 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 for details). This includes categories such as 'taxonomic' for *is-a* relations (e.g. *axe is_a tool*), 'function' for predicates denoting the use of an object (e.g. *hoe used_for_farming*), or again 'tactile' for features associated with the sense of touch (e.g. *toaster is_hot*). 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 1.3.2).[3]

[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,

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. But it also contradicts our hypothesis in §1.3.1 that quantification agreement might be linked to the English variety spoken by the coder: 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 conceptual rather than linguistic 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 does probably relate to actual disagreements in quantification, the other two seem to be artefacts of conceptual differences (what are 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 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

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

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 is highly dependent on their functional or sensory type. This indicates that generic information about concepts (e.g. there is a relation between cats and purring, or expressed in natural language: *Cats purr*) is more stable than model-theoretic knowledge (e.g. for most cats, it holds that there is a possible world where that cat purrs, *Most cats purr*). This finding corroborates the results of several studies that show that generics are acquired by children much earlier than quantifiers (e.g. Hollander et al., 2002). We think this is an important aspect to consider when using semantics as a potential cognitive representation. As far as model theory goes, it seems that converting a sentence into a logical form involving $\exists$ or $\forall$ may not always be cognitively straightforward. This calls for an underspecified formalisation of such sentences with actual, explicit quantification involving a further cognitive process – which may or may not be called upon by the speaker (see Herbelot and Copestake, 2011). Given the results reported here, it seems fair to assume that successful

communication relies on the generic, rather than the explicitly quantified level: a speaker is more efficient in uttering *tubas are loud* than the potentially controversial *some tubas are loud*. This would explain why so few sentences are prefixed with an explicit quantifier in English.

1.4 Conclusion

In this paper, we have presented an annotation layer for the McRae feature norms (McRae et al., 2005), which shows the natural language quantification 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. 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 attempted to match the used quantifiers to prevalence estimates. Under the assumption that more accurate estimates should result in better kappa agreement, we found that our annotators tended to conflate MOST and ALL, as well as NO and FEW. We hypothesised that when using their approximate number system, humans interpret MOST as a near-universal. We also showed that agreement is not correlated with the frequency 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. This led us to some remarks about the differences between conceptual information, as captured by feature norms, and the type of semantic representations assumed by model theory. Our data shows that while concepts reliably attract features that are relevant to them in production experiments, the resulting associations will in some cases correspond to very different set-theoretic interpretations across speakers.

So while overall, we observe good agreement on the quantification task (reaching κ_w^A=.59 for instances with a majority opinion), it seems unwarranted to assume that generalised quantifiers are always immediately cognitively available. We think this has consequences for the way we formalise quantification, but also for the types of models we develop for inference. While it seems fairly uncontroversial that *cats purr*, inferring that Sylvester the cat purrs is less than trivial. Even

less trivial is inferring the probable colour of a hypothetical bathtub: 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 hope, at any rate, that the dataset we are releasing will be of use when investigating such questions further.

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://disi.unitn.it/~bernardi/Papers/eacl12.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://www.cs.bham.ac.uk/~jxb/PUBS/BRM.pdf.

Clark, Robin and Murray Grossman. 2007. Number sense and quantifier interpretation. *Topoi* 26(1):51–62. ftp://babel.ling.upenn.edu/papers/faculty/robin_clark/topos.pdf.

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213. http://media.usm.maine.edu/~lenny/com375_office/weighted_kappa.pdf.

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://www.utsc.utoronto.ca/~gcree/pubs/CM03_JEPG.pdf.

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.eng.utah.edu/~cs6961/papers/hearst-coling92.pdf.

Herbelot, Aurelie and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*. Oxford, England, UK. http://www.aclweb.org/anthology/W11-0100#page=173.

Hollander, Michelle A, Susan A Gelman, and Jon Star. 2002. Children's interpretation of generic noun phrases. *Developmental Psychology* 38(6):883. http://gseacademic.harvard.edu/~starjo/papers/Hollander.pdf.

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. https://www.princeton.edu/~sjleslie/CogSci2009-inferences.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* pages 211–240. http://www.stat.cmu.edu/~cshalizi/350/2008/readings/Landauer-Dumais.pdf.

Lasersohn, Peter. 1999. Pragmatic halos. *Language* pages 522–551. http://semantics.uchicago.edu/kennedy/classes/s09/experimentalsemantics/lasersohn99.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://mentalmodels.princeton.edu/papers/ssk/ssk2011gog.pdf.

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.)